



中国人工智能产业发展联盟 AI 芯片技术选型目录 (2020 年)

中国人工智能产业发展联盟
计算架构与芯片推进组
2020 年 7 月

VISIT...

LANZAROTE
Caliente.COM

房中术(www.zuihaoziyuan.com)专注于男性增大增粗增长、阳痿、早泄

扫码免费领取资料



、壮阳、延时、强肾、回春、健身。女性缩阴、丰胸、减肥、化妆、瑜伽、保养、产后修复、盆底肌锻炼。两性健康,夫妻按摩,房中术,性姿势,性技巧,性知识等

扫码加微信lax98988领取资料



更多免费教程: 英语学习, 技能提升, PS 教学, 投资赚钱, 音乐教程, 口才教学, 情商提升, 风水教学, 心理学, 摄影知识, 幼儿教育, 书法学习, 记忆力提升等等.....

全站课程下载

课程不断增加

本站现资源容量已超 10T

入群联系 QQ: 167520299 或添加微信: 1131084518 (备注PDF)

阳痿早泄训练
皇室洗髓功视频教学
女人驻颜术
泡妞约炮万元课程
足疗养 SPA 教材
玉蛋功
马氏回春功
房中术张丰川
哲龙全套视频
增大盼你增大
国际男优训练
亚当德永早泄训练
洗髓功真人内部
皇室养生绝学道家洗髓功
【铁牛人会员课】男人必备技能, 理论讲解
实战高清视频
随意控制射精锻炼 视频+图片+文字
价值 1440 元第一性学名著<素女经房中养生
宝典视频>12 部
洗髓功修炼方法视频教学
陈见玉蛋功视频教学 女性缩阴锻炼
男性自然增大增长指南
强性健肾保健操 1-4
道家强肾系统锻炼功法
马氏回春功
12 堂课, 全面掌握男性健康问题 让你重燃自信
联系微信: 1131084518

- 1、东方性经
- 2、印度 17 式
- 3、口交技巧 3 部
- 4、港台性姿势 3 部
- 5、365 性姿势 6 部
- 6、泰国性爱密经 17 式
- 7、花花公子性技巧 6 部
- 8、阁楼艳星性技巧 7 部
- 9、古今鸳鸯秘谱全集 7 部
- 10、夫妻爱侣情趣瑜伽 2 部
- 11、古代宫廷性保健系列 14 部
- 12、汉唐宋元明清春宫图真人
- 13、柔软性爱宝典 日本 9800 课
- 14、李熙墨 3999 全套课
- 15、妖精性爱课 2888
- 16、李银河全套性课
- 17、领统统性课
- 18、德勇男性篇
- 19、德勇男性篇
- 20、缓慢性爱
- 21、亚当多体位搭配篇
- 22、亚当多体位结合篇
- 23、德勇克服早泄讲座练习
- 24、德勇以女性为中心得爱抚
- 25、加藤鹰接吻爱抚舌技
- 26、加藤鹰指技
- 27、加藤鹰四十八手入门
- 28、佐藤潮吹教学
- 29、佐藤男人体能锻炼+保健品介绍
- 30、佐藤男人早泄对抗训练
- 31、阿拉伯延时训练
32. 田渊正浩秘籍
33. 异性性快感集中训练教学
34. 自我愉悦锻炼密宗
35. 铁牛全套延时训练课
36. Pc 机锻炼真人视频教学
37. 印度性经全集 8 部
38. 21 世纪性爱指南
39. 香蕉大叔男女训练馆全套
40. 中美真人性治疗教学+理论
41. 女性闺房秘术
42. 幸福玛利亚性课
43. 陈见如何释放性魅力征服
44. 性爱技巧讲座全套
45. 性爱秘籍全套
46. 性爱误区讲座
47. 性病讲解大全
48. 性博士讲座合集
49. 性健康和性高潮合集
50. 性教育讲座合集
51. 性能力课堂合集
52. 性生活问题解析合集
53. 意外怀孕和避孕处理课堂
54. 性感地带探索
55. 性技巧讲座
56. 性健康与性卫生讲座
57. 性生活专家答疑
58. 性心理与性道德合集
59. 性爱宝典合集
60. 性爱技巧合集
61. 完美性爱演示
62. 完美性爱技术讲解
更多精品等你来解锁哦.....

每日免费获取报告

- 1、每日微信群内分享**7+**最新重磅报告；
- 2、每日分享当日**华尔街日报**、金融时报；
- 3、每周分享**经济学人**
- 4、行研报告均为公开版，权利归原作者所有，起点财经仅分发做内部学习。

扫一扫二维码

关注公众号

回复：**研究报告**

加入“起点财经”微信群。。





中国人工智能产业发展联盟 AI 芯片技术选型目录 (2020)

编写专家委员会

郑南宁	中国工程院院士
黄 如	中国科学院院士
楚 庆	紫光展锐首席执行官
陈云霁	中国科学院计算技术研究所所长助理、智能处理器研究中心主任
汪 玉	清华大学电子工程系主任、信息科学技术学院副院长

编写委员会主席

任鹏举	西安交通大学人工智能与机器人研究所副所长 / 副教授、博导
张蔚敏	中国信息通信研究院 工程师

编写成员单位 / 个人 (按名称的首字母来排序)

安谋科技 (中国) 有限公司
北京百度网讯科技有限公司
北京比特大陆科技有限公司
北京清微智能科技有限公司
成都启英泰伦科技有限公司
瑞芯微电子股份有限公司
高通信息科技 (上海) 有限公司
华为技术有限公司
深圳鲲云信息科技有限公司
赛灵思
上海海思技术有限公司
上海天数智芯半导体有限公司
上海深聪半导体有限责任公司
上海燧原科技有限公司
深圳云天励飞技术有限公司
颖脉信息技术 (上海) 有限公司
NVIDIA
紫光展锐
中科寒武纪科技股份有限公司

CONTENTS 目录

目录中产品顺序不分先后



一、引言	1
二、云端 / 数据中心	5
2.1 云端训练	6
2.1.1 北京百度网讯科技有限公司 – 百度昆仑 1 芯片	6
2.1.2 上海燧原科技有限公司 – 邃思通用人工智能训练芯片	7
2.1.3 NVIDIA–NVIDIA A100/V100/T4 Tensor Core GPUs	8
2.2 云端推断	10
2.2.1 赛灵思 – Alveo	10
2.2.2 中科寒武纪科技股份有限公司 – 寒武纪思元 100	14
2.2.3 中科寒武纪科技股份有限公司 – 寒武纪思元 270	15
2.2.4 北京比特大陆科技有限公司 – 比特大陆算丰 TPU 芯片 BM1684	16
2.2.5 深圳鲲鹏云信息科技有限公司 – 星空 x3 加速卡	18
2.2.6 华为技术有限公司 – 昇腾 310AI 处理器	20
2.3 小结	22
三、边缘计算	24
3.1 上海天数智芯半导体有限公司 – 天数智芯 Iluvatar CoreX I 芯片	25
3.2 NVIDIA–NVIDIA Jetson , NVIDIA T4 Tensor Core GPUs	27
3.3 紫光展锐 – 虎贲 T710	29
3.4 中科寒武纪科技股份有限公司 – 寒武纪思元 220 M.2	30
3.5 华为技术有限公司 – 昇腾 310AI 处理器	31
3.6 深圳鲲鹏云信息科技有限公司 – 星空 x3 加速卡	33
3.7 小结	36

四、终端	38
4.1 通用端侧产品	39
 4.1.1 NVIDIA–NVIDIA Jetson	39
 4.1.2 华为技术有限公司 – 昇腾 310AI 处理器	41
4.1.3 瑞芯微电子股份有限公司 – RK1808 AIoT 芯片方案	43
4.1.4 瑞芯微电子股份有限公司 – RK3399 Pro 人工智能端侧芯片方案	45
4.1.5 上海天数智芯半导体有限公司 – 天数智芯 Iluvatar CoreX I 芯片	47
4.2 智能驾驶	49
  4.2.1 赛灵思 – ZU7EV (Zynq® UltraScale+™ MPSoC)	49
 4.2.2 NVIDIA – NVIDIA DRIVE	53
4.3 安防、机器人、车载	54
4.3.1 深圳云天励飞技术有限公司 – 面向计算机视觉的深度学习神经网络处理器	54
4.3.2 紫光展锐 – 虎贲 T710	59
4.3.3 上海海思技术有限公司 – Hi3516D V300	60
4.3.4 上海海思技术有限公司 – Hi3519A V100	61
 4.3.5 上海海思技术有限公司 – Hi3559A V100	62
4.3.6 上海海思技术有限公司 – Hi3562 V100	63
4.3.7 上海海思技术有限公司 – Hi3566 V100	64
4.3.8 上海海思技术有限公司 – Hi3568 V100	65
4.3.9 上海海思技术有限公司 – Hi3569 V100	66
4.3.10 北京清微智能科技有限公司 – TX510	67
4.4 手机	68
4.4.1 华为技术有限公司 – Kirin 990 5G 处理器	68
4.4.2 高通信息科技（上海）有限公司 – 高通骁龙™ 865 移动平台	70
4.5 语音	72
  4.5.1 成都启英泰伦科技有限公司 – 人工智能语音识别降噪一体芯片 CI1103	72
 4.5.2 上海深聪半导体有限责任公司 – 人工智能语音芯片	75

4.5.3 北京清微智能科技有限公司 – TX210	77
4.6 小结	78
五、IP 技术	80
5.1 安谋科技（中国）有限公司 – 周易 AIPU	81
 5.2 颖脉信息技术（上海）有限公司 – PowerVR NNA 神经网络加速器系列 IP	83
5.3 小结	86
附录	88

说明



图 1 产品验证测试 logo

AI 芯片技术前有图 1 logo，代表企业在报名提交产品文本资料后，同时提供产品技术测试报告并参与由 AIIA 委托中国信息通信研究院进行的验证测试，但具体实测性能不公开。



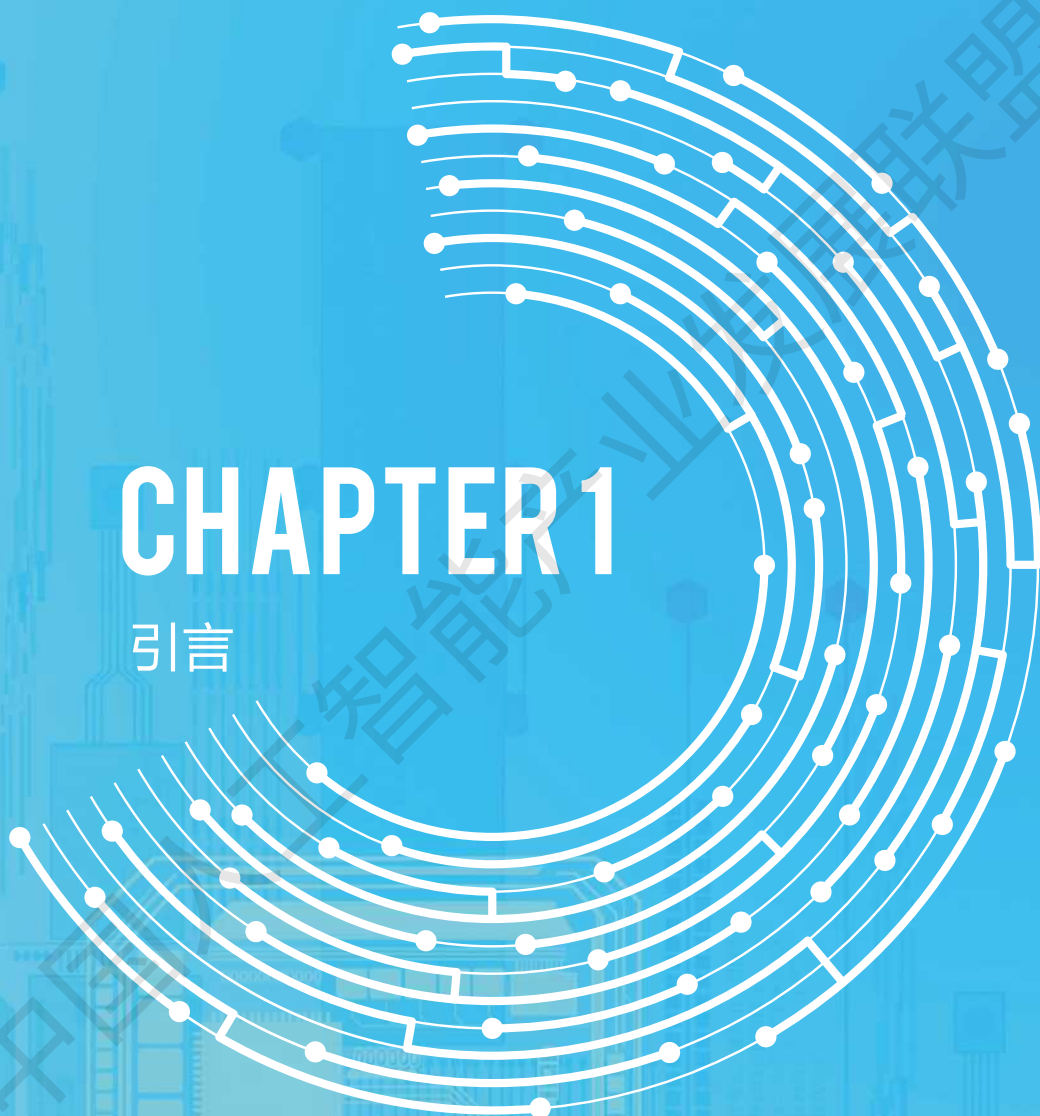
图 2 AIIA DNN benchmark 项目 logo

AI 芯片技术前有图 2 logo，代表在企业提交产品资料介绍时，同时提交产品技术测试报告，并参与 AIIA DNN benchmark 项目测试，具体实测性能在 AIIA DNN benchmark 官网可查看（<http://www.aiiaorg.cn/benchmark>）。

声明：本选型目录的编写工作以企业自愿报名方式开展，所有相关产品的数据文本资料均由企业提供且经企业审核，如因内容不当造成直接或间接损失，AIIA 不承担任何责任。

CHAPTER 1

引言



背景与意义

当前随着人工智能理论和技术的日益成熟，应用范围不断扩大，目前已广泛应用于计算机科学、金融贸易、医药、诊断、重工业、运输、远程通讯、在线和电话服务、法律、科学发现、玩具和游戏、音乐等诸多方面。

算力作为承载人工智能应用的平台和基础，其发展推动了整个人工智能系统的进步和快速演进，是人工智能最核心的要素之一。以人工智能应用为主要任务的，面向智能计算的处理器的相关设计方法与技术已成为国内外工业界和学术界共同角逐的热点，国内外企业纷纷布局 AI 芯片。

AI 芯片的应用场景不再局限于云端，部署于智能手机、安防摄像头、及自动驾驶汽车等终端的各项产品日趋丰富。除了追求性能提升外，AI 芯片也逐渐专注于特殊场景的优化。

为了进一步促进供需对接，为 AI 芯片供应商和需求商提供交流的平台，中国人工智能产业发展联盟（以下简称“AIIA”或联盟）计算架构与芯片推进组启动“AI 芯片技术选型目录”（以下简称“选型目录”）的工作，希望为 AI 芯片的可持续发展，服务和促进相关产业发展壮大贡献一份力量。

内容与目的

选型目录在 1 年前开始面向全球企业征集，内容包含 AI 芯片产品的：

- 技术概述：主要介绍 AI 芯片产品的技术特点与应用场景及功能概述；
- 技术指标：主要介绍 AI 芯片产品实现的技术及其特点；
- 应用案例：主要介绍该款 AI 芯片产品已经应用的场景与经典的案例介绍；
- 产品特征 / 技术特点：详细介绍 AI 芯片产品的各项指标或特定应用领域的技术描述；
- 专利状态：作为选填项，披露产品专利状态
- 产品信息：以列表形式，展示 AI 芯片的主要产品信息，包括功耗、芯片特征、上市时间、深度学习计算能力及工艺等。

征集 AI 芯片产品信息后，由 AIIA 委托中国信息通信研究院进行芯片验证测试。参与产品验证测试的 AI 芯片，将以“tested by AIIA” logo 体现在产品介绍页，如图 1 所示，代表该 AI 芯片有真实产品但真实性不能公开。



图 1 产品验证测试 logo

为了帮助企业、系统集成厂商、开发者等准确评价芯片的技术水平和在国际上行业内的地位，AIIA 于 2017 年发起的 AIIA DNN benchmark 也为 AI 芯片选型目录工作提供技术支撑。为了进一步帮助企业产品宣传，帮助用户了解 AI 芯片的真实性能表现，参与 AIIA DNN benchmark 的 AI 芯片，将以“AIIA DNN benchmark” logo 体现在产品介绍页，如图 2 所示，代表该芯片可从 AIIA DNN benchmark 官网查到测试数据（<http://www.aiiaorg.cn/benchmark>）。



图 2 AIIA DNN benchmark 项目 logo

经过近半年的筹备工作，推进组完成材料汇总，测试验证并组织行业内专家对目录内容进行专家评审工作。AIIA 2020 年度首批“AI 芯片技术选型目录”，共收到 22 家企业报名，反馈 49 款 AI 芯片相关产品技术，经过测试、评审首批目录包含 19 家企业共 46 款 AI 芯片产品技术。

人工智能芯片按照目前的应用场景来看，主要分为训练和推断两类场景，按部署位置又可分为云端、边缘和终端。AIIA “AI 芯片技术选型目录”具体根据应用场景与部署位置，包含云端训练、基于云端、边缘和终端推断的四大产品形态。又依据行业应用，AI 芯片技术选型目录根据垂直应用场景如图 3 所示给出。



图 3 AI 芯片根据行业应用分类图

AIIA 希望通过“AI 芯片选型目录”的工作，一方面提供选型参考，希望收入 AI 芯片目录的产品信息可以作为用户选型依据之一；另一方面是帮助芯片企业宣传产品，同时通过测试验证以及依托 AIIA DNN benchmark 项目，来切实反映当前 AI 芯片的真实性能。

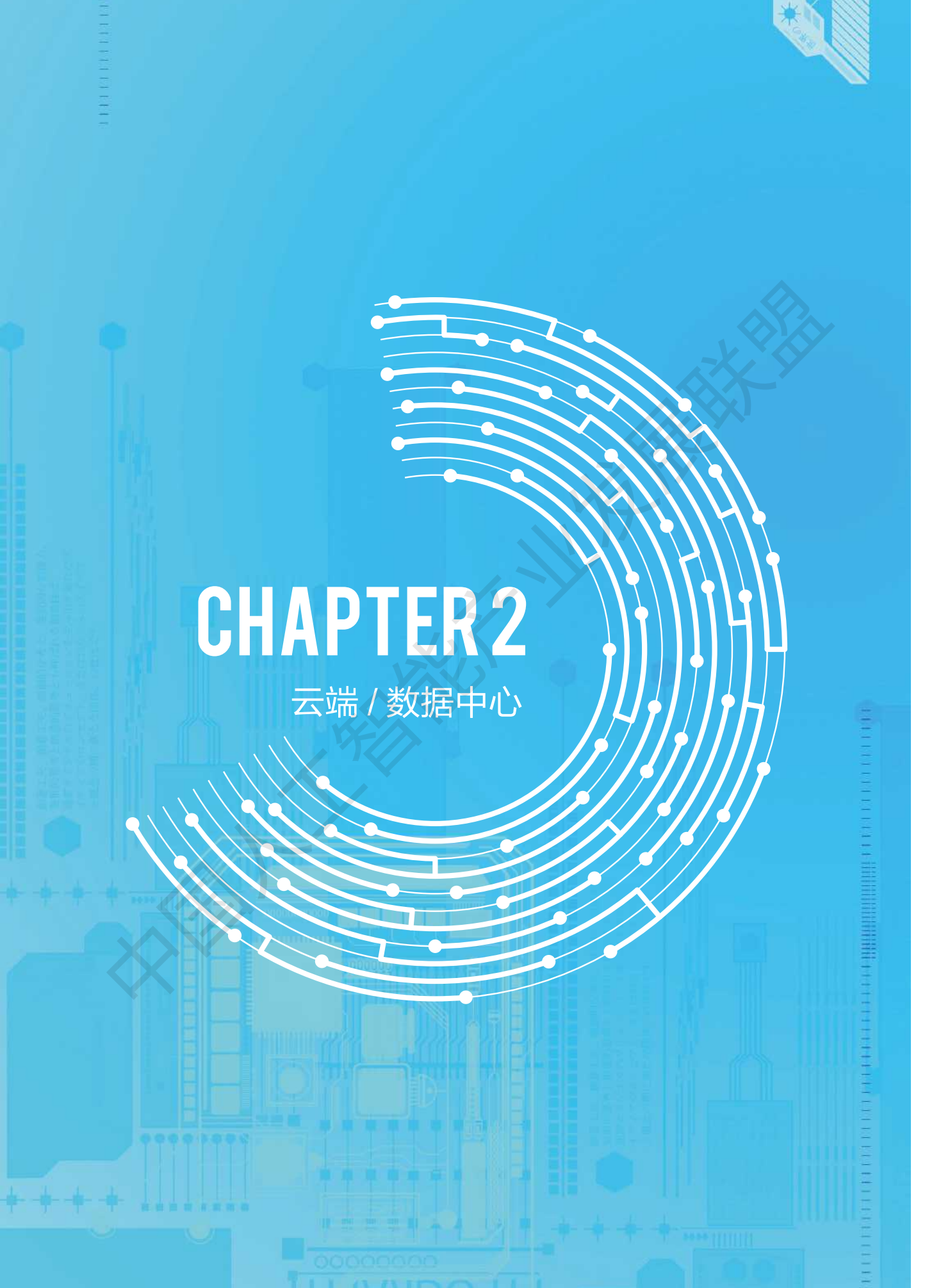
后续技术选型目录将以联盟内 AI 芯片供应商和需求商的真实案例为切入点，继续迭代更新目录内容，同时推动符合用户需求的测试评估，以促进产业健康发展，同时助力人工智能生态建设。



中国人工智能产业发展联盟

CHAPTER 2

云端 / 数据中心





2.1 云端训练

2.1.1 北京百度网讯科技有限公司 - 百度昆仑 1 芯片



【技术概述】

百度昆仑 1 芯片面向 AI 通用计算，基于百度 XPU 架构，既可以用于训练，也可以用于推理，能全面支持语音，图像，自然语言处理等应用。

【技术指标】

百度昆仑 1 芯片基于百度 XPU 架构，采用三星 14nm 工艺，HBM2 和 PCIe4.0 等技术，给用户高性能、高带宽、低功耗、高灵活性、高度可编程等优势。

【产品信息】

【应用案例】

百度 ERNIE 模型，应用于翻译、搜索排序等场景。百度网盘里的图像模型；百度云主机；度云的工业质检客户。

【产品特征 / 技术特点】

1. 高性能，峰值 256Tops；
2. 高带宽，片外内存带宽 512GB/s；
3. 高度可编程性，能提供 C/C++ 的编程，用户可以自由开发自己的算子。

【专利状态】（未提供）

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
百度昆仑 1	150W	16GB	512GB/s	2019 年 11 月	256Tops	64Tops	16Tops	14nm



2.1.2 上海燧原科技有限公司 - 邃思通用人工智能训练芯片

【技术概述】

邃思芯片集成神经元处理器架构和数据处理引擎，通过可编程的通用张量 / 向量运算核心，支持各类数据精度的主流深度学习训练负载。

【技术指标】

邃思芯片采用 12nm FinFET 工艺，2.5D 封装，总计 141 亿个晶体管。其宣称单芯片提供 20TFLOPS@FP32 及 80TFLOPS@BF16/FP16 的算力，最大功耗 190W。

邃思芯片主要技术特点包括：

- 自主指令集的神经元处理器（SIP），可编程的通用张量 / 向量运算核心，支持张量 / 向量 / 标量计算；

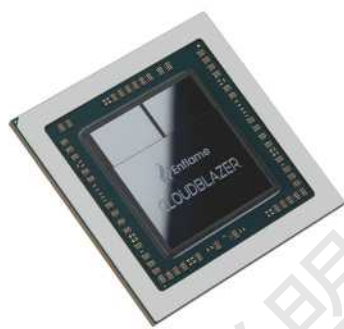
- 神经元处理集群（SIC），由 8 个神经元处理器构成，4MB 共享集群缓存；

- 自主研发 2.5D 封装，集成 16GB HBM 存储，提供 512GB/s 带宽；

- 自主 ESL（燧原智能互联，Enflame Smart Link）片间互联引擎，每通道双向 50GB/s，4 通道总共 200GB/s 接口带宽，通道延时小于 1μs；

- 系统稳定性控制，支持服务器级别 RAS 需求（Reliability, Availability and Serviceability, RAS）。

- 支持 CNN、RNN、LSTM、BERT 等网络模型



- 支持数据类型：FP32/FP16/BF16/INT8/INT16/INT32 等

【应用案例】（未提供）

【产品特征 / 技术特点】

- 云燧 T10 是基于邃思芯片打造的人工智能训练加速卡，20TFLOPS@FP32，最大功耗 225W，支持 FP32/FP16/BF16/INT8/INT16/INT32 等多种数据类型；支持 PCIe 4.0 接口，提供 64GB 带宽以及 200GB/s 片间互联带宽；

- 驭算计算和编程平台主要包含分布式调度、软件开发包（SDK）和设备驱动层，用户可以将已有训练系统，无缝迁移至邃思平台上运行，无需二次开发或者模型转换工作。

- 云燧 T10 人工智能训练加速卡面向云端数据中心，可广泛应用于互联网、金融、教育、医疗、工业及政务等人工智能训练场景。

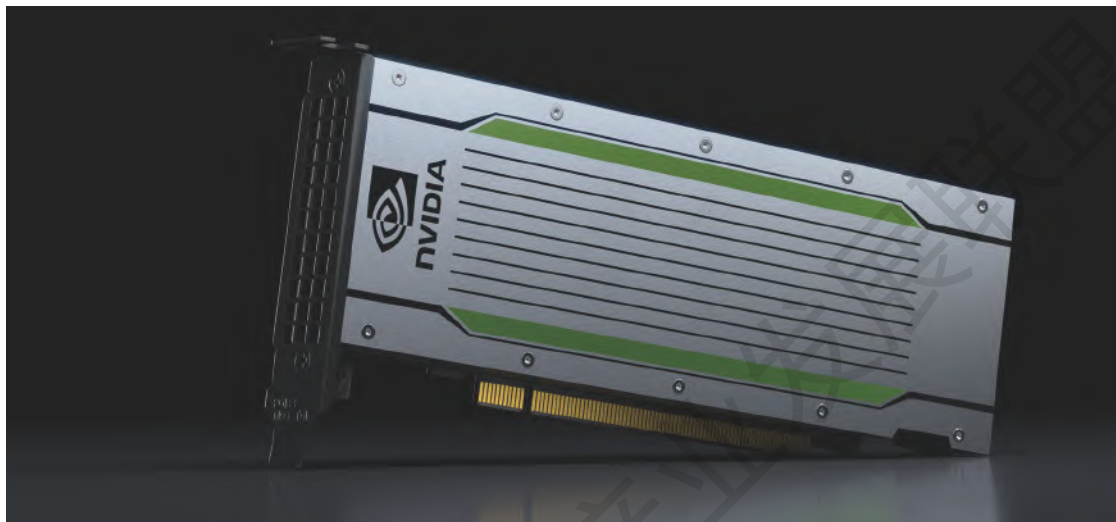
【专利状态】

燧原科技及其子公司目前已获得 24 项授权专利，其中包括 19 项发明专利，另有近 20 件专利正在申请进程中。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
邃思	190W（最大功耗）	16GB HBM	512GB /s	2020 年 6 月	80TOPS	80TFLOPS	20TFLOPS	GF 12nm FinFET

2.1.3 NVIDIA-NVIDIA A100/V100/T4 Tensor Core GPUs



【技术概述】

为数据中心 GPU，可加快 AI、高性能计算 (HPC) 和图形技术的发展。

【技术指标】

NVIDIA Ampere GPU 架构配备 540 亿个晶体管，为 7 纳米芯片架构，集训练和推理于一身，具有多实例 GPU、TF32、结构化稀疏等特性。

NVIDIA Volta GPU 架构配备 640 个 Tensor 内核，可提供每秒超过 100 万亿次 (TFLOPS) 的深度学习性能，是上一代 NVIDIA Pascal 架构的 5 倍以上。

NVIDIA Turing GPU 架构支持实时光线追踪、AI、模拟和光栅化技术。

【应用案例】

利用 NVIDIA A100/V100/T4 GPU 可以更快地处理要求最严格的高性能计算 (HPC) 和超

大规模数据中心工作负载。现在，数据科学家和研究人员可以在能源勘探和深度学习等应用场合中解析 PB 级的数据，速度比使用传统 CPU 快几个数量级。NVIDIA A100/V100/T4 加速器可以超快速度运行更大型的模拟。此外，NVIDIA GPU 还能作为虚拟桌面、应用程序和 workstation 提供超高性能和用户密度。

【产品特征 / 技术特点】

NVIDIA A100 采用了 NVIDIA Ampere 架构的突破性设计，集 AI 训练和推理于一身，宣称其性能相比于前代产品提升了 20 倍。作为一款通用型工作负载加速器，A100 还被设计用于数据分析、科学计算和云图形。

NVIDIA V100 运用 NVIDIA Volta 架构，拥有 640 个 Tensor 内核，是世界上第一个突破 100 万亿次 (TFLOPS) 深度学习性能障碍的



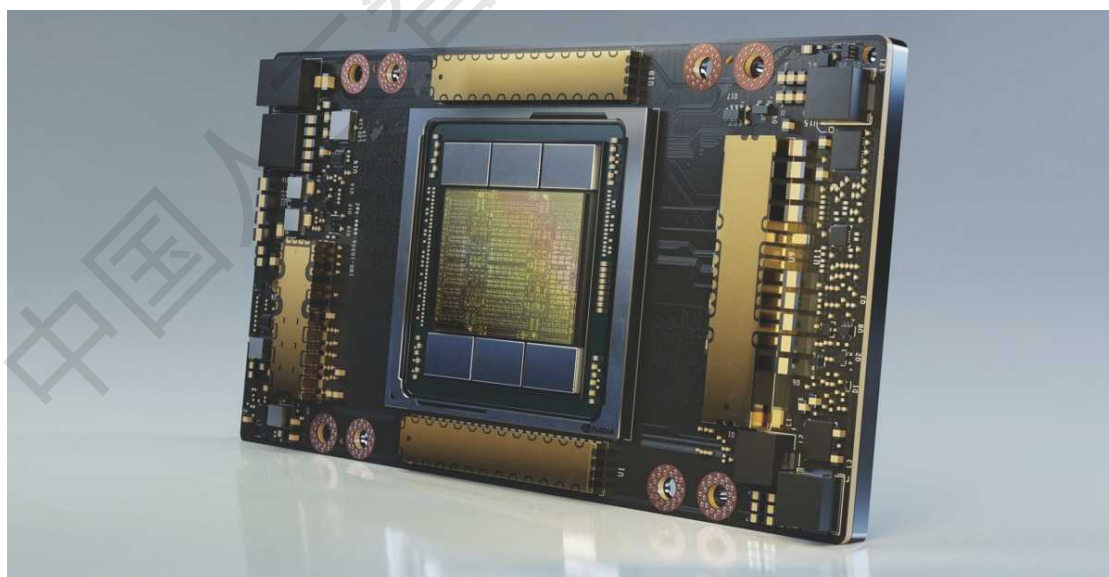
GPU。

NVIDIA T4 GPU 推理加速器搭载 NVIDIA Turing Tensor Core，提供多精度推理性能，以加速人工智能的各种应用。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	显存	带宽	上市时间	深度学习计算能力					工艺
					INT8	FP32	FP64	TF32	FP64 Tensor Core	
NVIDIA A100	400w	40GB HBM2	1555 GB/s	2020 年 5 月	624 TOPS 1248 TOPS* (* 采用稀疏技术)	19.5 TFLOPS	9.7 TFLOPS	312TFLOPS (* 采用稀疏技术)	19.5 TFLOPS	7nm
NVIDIA V100 (NVLink)	300w	32/16 GB HBM2	900 GB/s	2017 年 5 月		15.7 TFLOPS	7.8 TFLOPS			12nm
NVIDIA V100 (PCIe)	250w	32/16 GB HBM2	900 GB/s	2017 年 5 月		14 TFLOPS	7 TFLOPS			12nm
NVIDIA V100S	250w	32 GB HBM2	1134 GB/s	2019 年 11 月		16.4 TFLOPS	8.2 TFLOPS			12nm
NVIDIA T4	70w	16 GB GDDR6	320+ GB/s	2018 年 3 月	130TOPS	8.1 TFLOPS				12nm





2.2 云端推断

2.2.1 赛灵思 – Alveo



【技术概述】

赛灵思 Alveo™ 数据中心加速器卡专为现代数据中心多样的应用需求而设计。Alveo 加速器卡可通过对加速资源的重新配置，适应持续算法优化，在降低总成本的同时，灵活支持各类工作负载。

赛灵思推出的 Vitis 统一软件平台为各类软件和 AI 推理应用开发提供统一编程模型，帮助用户加速实现从 C/C++、Python、Caffe、Tensorflow 到差异化应用落地的开发过程。

【技术指标】

赛灵思 Alveo™ 数据中心加速器卡基于 Xilinx 16nm UltraScale™ 架构，使用赛灵思堆叠硅片互联 (SSI) 技术来实现 FPGA 容量、带宽和功耗效率，通过结合多个超逻辑区域 (SLR) 来增大密度。Alveo 加速卡旨在加速服务器或工作站中的机器学习、数据分析和视频处理等计算密集型应用。

AlveoU50 卡采用 XCU50 FPGA 包括 2 个 SLR，配备 PCIe Gen4 和 8G HBM2，每秒 100G 网络连接，以高效能 75 瓦、小尺寸形式为金融计算、机器学习、计算存储以及数据搜索与分析工作负载提供优化加速。

Alveo U200 卡采用 XCU200 FPGA 包括 3 个 SLR，Alveo U250 卡使用 XCU250 FPGA 包括 4 个 SLR。二者均可连接到 PCI Express® 的 16 个通道，最高运行速度 8GT/s (Gen3)，也可以连接到 4 根 DDR4 16 GB 2400 MT/s 64 位含纠错码 (ECC) 的 DIMM，总计 64 GB 的 DDR4。

Alveo U280 加速卡采用 XCU280 包括三个 SLR，底部 SLR (SLR0) 集成一个 HBM 控制器，与相邻的 8 GB HBM2 内存接口连接。底部 SLR 还连接到 PCI Express® 的 16 个通道，这些通道可以最高 16 GT/s (Gen4) 的速度运行。SLR0 和 SLR1 都连接到 DDR4 16 GB 2400 MT/s 64

位含纠错码 (ECC) 的 DIMM, 总计 32 GB 的 DDR4。

【应用案例】

Alveo 加速卡已经开发和认证的关键应用涵盖 AI/ML、视频转码、数据分析、金融风险建模、安全和基因组学等。Algo-Logic Systems Inc、Bigstream、BlackLynx Inc.、CTAccel、Falcon Computing、Maxeler Technologies、Mipsology、NGCodec、Skreens、SumUp Analytics、Titan IC、Vitesse Data、VYUsync 和 Xelera Technologies 等合作伙伴开发完成的应用已经投入部署。此外, Dell EMC、Fujitsu Limited 和 IBM 等顶级 OEM 厂商也在同赛灵思合作, 认证采用 Alveo 加速器卡的多个服务器 SKU。

一, Alveo-Mipsology, Zebra 搭载 Alveo U50 数据中心加速卡用于神经网络推理, 板卡功耗 75W, 关键功能:

快速卷积神经网络 (CNN) 推理——Zebra 设立了新的性能标准;

支持所有神经网络——Zebra 可加速任何 CNN 的所有层级;

易于使用——Zebra 是“即插即用”的解决方案, 无需对神经网络进行修改或再训练;

不改变软件环境——无新语言、框架或工具。Zebra 在 TensorFlow、PyTorch、ONNX、Caffe 和 MXNet 模型内运行;

可扩展性、灵活性和适应性——Zebra 可替代 GPU 或在任何位置补充 CPU。

二, Alveo-SK 电讯基于人工智能的实时物理入侵和盗窃检测服务, 基于赛灵思 Alveo U250 加速器卡, 可提供充足的吞吐量与精准度。

【产品特征 / 技术特点】

Alveo U50 数据中心加速器卡采用赛灵思 UltraScale+™ 架构, 使用半高半长的外形尺寸和低于 75 瓦的低包络功耗。该卡支持第四代 PCIe 及高带宽存储器 (HBM2), 每秒 100G 网络连接, 面向各种类型的服务器部署。

Alveo U50 加速器卡产品详细介绍:

规格	U50
产品 SKU	A-U50-P00G-PQ-G
卡总耗电量 ¹	75W
散热解决方案	被动
重量	300g - 325g
外形尺寸	半高半长
网络接口	1x QSFP28
PCIe 接口 ^{2,3}	Gen3 x16, Gen4 x8, CCIX
HBM2 总容量	8 GB
HBM2 带宽	201 GB/s ⁴
查找表 (LUTs)	872K
寄存器	1,743K
DSP slices	5,952
最高分布 . RAM	24.6 Mb
36 Kb 块 RAM	1344 (47.3 Mb)
288 Kb UltraRAM	640 (180.0 Mb)
GTY 收发器	20
符合部署条件	有

注释:

1. Alveo U50 卡为 FPGA 结构与 HBM 存储器分别搭配了单独的电源供电线。开发者必须确保他们的设计不会从每条供电线汲取过多电源。如需了解更多信息, 请参阅《Alveo U50 数据中心加速器卡安装指南》(UG1370)。

2. PCIe 接口可以配置为支持各种链接宽度和速度。最高为 Gen3 (8 Gt/s) x16、Gen4 (16 Gt/s) x8 或以 16 GT/s x8 运行的 CCIX。PCIe 接口还可以配置为双 x8 接口, 并连接到支持 PCIe 分叉的主机。

3. 此块在兼容模式下以 16.0 Gt/s (Gen4) 的速度运行。请参阅《UltraScale+ Devices Integrated Block for PCI Express LogiCORE IP 产品指南》(PG213) 了解兼容模式的详情。



4. 通过 PCIe 3.3V 电源供电线提供的 HBM2 功耗仅限为 10W。使用 HBM2 可实现的性能受限于此功耗限制，并且因设计而异。HBM2 的额定带宽为 201 GB/s。在不兼容 PCIe 的规格中，针对 A-U50DD-P00G-ES3-G 和 A-U50-P00G-PQ-G 卡测得的 HBM2 带宽峰值为 316 GB/s。

Alveo U200 和 U250 数据中心加速器卡采

用 PCIe Gen3 x16，设计用于加速高计算强度应用，如机器学习、数据分析和视频处理。

注释：

1. 225W PCIe CEM 卡通过标准连接器 12V 电源获取 65W，通过 AUX 连接器 12V 电源获取另外 150W。此卡上不使用标准连接器提供的 3.3V 电源。CEM 卡需将 150W PCIe AUX 电源线缆连接到本卡。

Alveo U200/U250 加速器卡产品详细介绍

规格	U200 U250		U200 U250	
	动散热版	被动散热版	主动散热版	被动散热版
产品 SKU	A-U200-A64G-PQ-G	A-U200-P64G-PQ-G	A-U250-A64G-PQ-G	A-U250-P64G-PQ-G
散热解决方案	主动	被动	主动	被动
	1122g	1066g	1122g	1066g
外形尺寸	全高、全长、双宽度	全高、¾ 长、双宽度	全高、全长、双宽度	全高、¾ 长、双宽度
卡总耗电量 ¹	225W		225W	
网络接口	2x QSFP28		2x QSFP28	
PCIe 接口	Gen3 x16		Gen3 x16	
查找表 (LUT)	1,182K		1,728K	
寄存器	2,364K		3,456K	
DSP slice	6,840		12,288	
UltraRAM	960		1,280	
DDR 总容量	64 GB		64 GB DDR	
DDR 最大数据率	2400 MT/s		2400 MT/s	
DDR 总带宽	77 GB/s		77 GB/s	

Alveo U280 数据中心加速器卡专为计算和存储工作负载而设计，拥有 8GB HBM2 + 32GB DDR4 内存、1.1M LUT、8.5k DSP 片、每秒 100G 双网络连接，并支持第四代 PCIe 和 CCIX 互联标准。

Alveo U280 加速器卡产品详细介绍：

规格 ¹	主动散热版	被动散热版
产品 SKU	A-U280-A32G-DEV-G ³	A-U280-P32G-PQ-G ³
卡总耗电量 ²	225W	225W
散热解决方案	主动	被动
重量	1187g	1130g
外形尺寸	全高、全长、双宽度	全高、¾ 长、双宽度
网络接口	2x QSFP28	
PCIe 接口 ^{4,5}	Gen3 x16, Gen4 x8, CCIX	
HBM2 总容量	8 GB	
HBM2 带宽	460 GB/s	
查找表 (LUTs)	1,079K	
寄存器	2,607K	
DSP slices	9,024	
UltraRAMs	960	
DDR 总容量	32 GB	
DDR 最大数据率	2400 MT/s	
DDR 总带宽	38 GB/s	

注释：

1. 规格适用于量产 U280 卡。U280 ES1 卡规格和功能的任何差异都记录在 U280 ES1 已知问题赛灵思答复记录 71975 中。

2. 225W PCIe CEM 卡通过标准连接器 12V 电源获取 65W，通过 AUX 连接器 12V 电源获取另外 150W。此卡上不使用标准连接器提供的 3.3V 电源。CEM 卡需将 150W PCIe AUX 电源线缆连接到本卡。

3. 被动散热卡和主动散热卡都符合部署条件。主动 DEV 器件编号包含一条 USB 线用于开发。如需了解有关不含 USB 线的主动卡的批量订购选项的

信息，请联系您当地的赛灵思销售代表。

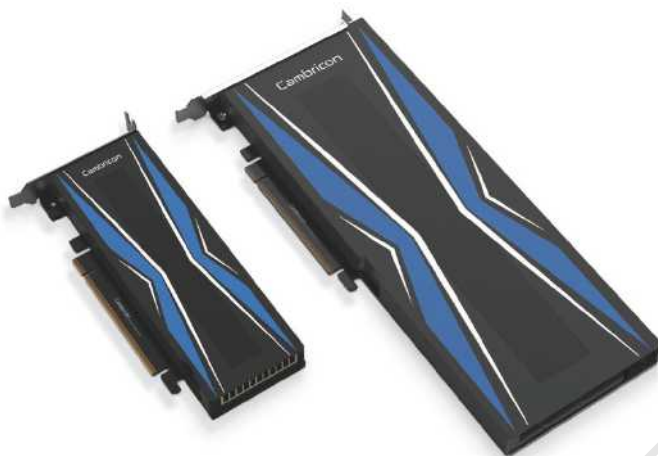
4. PCIe 接口可以配置为支持各种链接宽度和速度。最高为 Gen3 (8 Gt/s) x16、Gen4 (16 Gt/s) x8 或以 16 GT/s x8 运行的 CCIX。PCIe 接口还可以配置为双 x8 接口，并连接到支持 PCIe 分叉的主机。

5. 此块在兼容模式下以 16.0 Gt/s (Gen4) 的速度运行。请参阅《UltraScale+ Devices Integrated Block for PCI Express LogiCORE IP 产品指南》(PG213) 了解兼容模式的详情。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Alveo U50 (VU35P)	50W	8GB	316GB/s	2019 年 8 月	16.2	/	/	16nm UltraScale™
Alveo U200 (VU9P)	100W	64GB	77GB/s	2018 年 10 月	18.6	/	/	
Alveo U250 (VU13P)	110W	64GB	77GB/s	2018 年 10 月	33.3	/	/	
Alveo U280 (VU37P)	100W	/	/	2018 年 11 月	24.5	/	/	



2.2.2 中科寒武纪科技股份有限公司 - 寒武纪思元 100



【技术概述】

为云端推理提供运算能力支撑。

【技术指标】

INT8 算力 32TOPS, 内置硬件编解码引擎

【应用案例】

计算机视觉、语音识别、自然语言处理。

【产品特征 / 技术特点】

1. 通用智能，支持计算机视觉、语音识别、自然语言处理等多模态智能处理；
2. 针对深度学习定制的指令集和处理器架构，具有更优的能效比；
3. 完善软件开发环境 NeuWare，包括应用开发、功能调试、性能调优等。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
思元 100	75w	8GB/16GB	102GB/s	2018 年 05 月	32TOPS	16TFLOPS	/	16nm



2.2.3 中科寒武纪科技股份有限公司 – 寒武纪思元 270



【技术概述】

为高效能 AI 推理设计的数据中心级 PCIe 智能加速卡。

【技术指标】

支持多种精度，比上一代加速芯片计算能力提高 4 倍，INT8 算力 128TOPS。

【应用案例】

广泛支持视觉、语音、自然语言处理以及传统机器学习等高度多样化的人工智能应用，帮助 AI 推理平台实现高效能比。

【产品特征 / 技术特点】

1. 支持 INT16、INT8、INT4、FP32、FP16 多种精度；
2. 内置视频和图片编解码器，有效降低 CPU 前处理负载和 PCIe 带宽占用；
3. 计算弹性，支持多类神经网络，寒武纪 Neuware 软件栈部署推理环境；
4. 可编程，基于 Bang 语言编程环境可对计算资源定制，满足多样化需求。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
思元 270	70w	16GB	102GB/s	2019 年 06 月	128 TOPS	/	/	16nm

2.2.4 北京比特大陆科技有限公司 – 比特大陆算丰 TPU 芯片 BM1684



【技术概述】

BM1684 是比特大陆面向深度学习领域自主研发的第三代张量处理器（TPU），是聚焦视频图像分析的云端及边缘的人工智能推理芯片。

【技术指标】

1. 芯片：BM1684 聚焦视频图像分析，是云端及边缘的人工智能推理芯片；
2. AI 算力：17.6TOPS INT8，Winograd 卷积加速下最高可达 35.2TOPS，实测推理性能较上一代提升约 5 倍以上；
3. AI 架构：本芯片基于自主研发的 TPU 架构；
4. 典型功耗：16W；
5. 视频解码：支持 H264 / H265 解码，最大分辨率 8192x8192，支持 4K/8K。H264 和 H265 解码都支持 32 路高清 30FPS @1080P，可处理数十路视频智能分析全流程；
6. 图像解码：支持 JPEG 解码和编码，均可支持 480 张 / 秒 @1080P；

7. CPU：八核 A53，主频 2.3GHz；

8. 内存：LPDDR4X，带宽 68.3GB/s；

9. AI 框架支持：Caffe, Tensorflow, PyTorch, MXNet, PaddlePaddle 飞桨等；

【应用案例】

1. 北京海淀城市大脑中的 AI 计算推理，百度大脑 Paddle TPU 计算集群，处理千路级别的高清视频结构化。
2. 福州长乐区的 AI 算力中心 TPU 计算集群，处理上千路的高清视频结构化、人脸识别等。
3. 福州城市大脑的 AI 计算推理 TPU 计算集群，处理千路级别的高清视频结构化。
4. 优必达（Ubitus）的互联网云端游戏 AI 加速 TPU 加速游戏中的图像特征提取，画面增强，超分辨率等。
5. 某新零售客户

新零售货柜的边缘计算盒子，可进行商品自动识别。

【产品特征 / 技术特点】

1. BM1684 芯片技术特点是：
 - 1) TPU 芯片架构自主研发，相关专利申请达到 270 项以上；
 - 2) 性能功耗比高，在 16W 情况下，最高可达到 35.2T 性能（Winograd 加速）；
 - 3) 视频解码路数多，支持 32 路 H264 和 H265 高清 30FPS @1080P 硬解码；
 - 4) 视频全流程处理能力强，可达到 16~32 路典型视频结构化 / 人脸分析路数；
 - 5) AI 工具链完备，Caffe, Tensorflow, PyTorch, MXNet, PaddlePaddle 都支持；
 - 6) 部署场景灵活，云端和边缘均可部署；
 - 7) 使用灵活，可工作于 PCIE 从设备模式或者 SOC 主设备模式；
2. 板卡产品特色：
 - 1) 75W 半高半长 PCIE 小卡，峰值算力 105.6T，无需外接供电；
 - 2) 宣称其性能赶上并部分超越英伟达推

理 GPU T4；

- 3) 视频解码路数多，96 路高清视频 30FPS 解码；

3. 模组产品特色：

- 1) AI 边缘计算模组，17.6T/35.2T；
 - 2) 视频解码支持 32 路 H264 和 H265 高清 30FPS @1080P 硬解码；
 - 3) 视频全流程处理可达到 16~32 路典型视频结构化 / 人脸分析；
- ### 4. 盒子产品特色：
- 1) AI 边缘计算盒子，17.6T/35.2T；
 - 2) 视频全流程处理能力可达到 16~32 路典型视频结构化 / 人脸分析；
 - 3) 可室外部署，无风扇散热，-20℃至 +60℃宽温；
 - 4) 接口支持 USB/HDMI/RS-485/RS-232 等；

【专利状态】

比特大陆自主研发的 TPU 芯片，自主研发架构，相关专利申请达到 270 项以上；

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
BM1684	16W	12GB LPDDR4X	68.3GB/s	2019 年 9 月	17.6T (普通模式); 35.2T (Winograd 加速模式)		2.2T	12 nm

2.2.5 深圳鲲云信息科技有限公司

- 星空 x3 加速卡



Corerain
鲲云科技

星空 X3 加速卡

【技术概述】

星空 X3 加速卡为鲲云推出的面向边缘端和数据中心进行深度学习推断的 AI 计算加速卡，搭载鲲云自研的定制数据流 CAISA 芯片，采用无指令集的架构方式，为支持深度学习的边缘和数据中心服务器提供计算加速方案。

【技术指标】

1. 支持 ResNet、VGG、YOLO 等多个主流 CNN 算法模型
2. 提供 RainBuilder 编译工具链，支持端到端算法开发和部署
3. 支持 TensorFlow、Caffe、PyTorch 及 ONNX (MXNet) 等主流深度学习框架开发的算法模型

4. 其宣称芯片利用率可达 95.4%

【应用案例】

1. 工业领域：鲲云与合作方针对某电容上激光刻蚀字符推出基于深度学习的 OCR 识别方案，识别算法采用了深度学习方法，可识别字母、数字以及数十种特殊字符，字符识别精度可达到 99.99%。采用搭载鲲云 CAISA 芯片的星空加速卡，识别延迟可以降低到 50ms，识别速度可达到 20pcs/s。

2. 电力领域：

鲲云与合作方通过研究适用于电力无人机巡检的目标检测深度学习算法，实现基于人工智能技术的电力无人机智能巡检、数据采集，并在服务器端进行高质量的图片数据分析，提高巡检效率，降低

巡检工作量。

3 支持分类、目标检测以及语义分割类深度学习算法；

【产品特征 / 技术特点】

1 其宣称芯片利用率可达 95.4%；

4 Batch size 不敏感；

2 时延：3ms 分类延时；

5 温度范围：-20℃ ~70℃

6 实测 Benchmark：

算法网络	ResNet-50	ResNet-152	YOLOv3	SSD-ResNet50	U-Net Industrial
延时	3.05 ms	8.68 ms	31.06 ms	20.8 ms	74.07 ms
吞吐率	1306.93 FPS	460.27 FPS	125.75 FPS	182.16 FPS	54.01 FPS
芯片利用率	92.32%	95.43%	82.37%	77.06%	64.97%

注：Batch=4，INT 8

【专利状态】

已申请专利 110 件，包括 109 件发明专利、1 件外观设计专利；

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
星空加速卡 x3	峰值 56W	两组 DDR4 SDRAM, 8GB	32Gbps	2020 年 6 月	10.9 TOPS	N/A	N/A	28nm

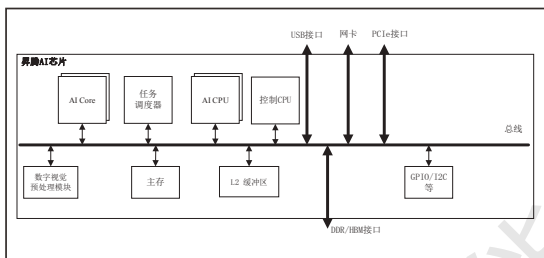
2.2.6 华为技术有限公司 – 昇腾 310AI 处理器



【技术概述】

昇腾 AI 处理器的主要架构组成：

- 芯片系统控制 CPU（Control CPU）
- AI 计算引擎（包括 AI Core 和 AI CPU）
- 多层级的片上系统缓存（Cache）或缓冲区（Buffer）
- 数字视觉预处理模块（Digital Vision Pre-Processing, DVPP）等



标检测、光学字符识别、视频处理和自然语言处理领域的各个模型，均可以在昇腾 310 处理器上得到很好的技术支持。

【技术指标】

规格	描述
架构	达芬奇
性能	Up to 8T @FP16 Up to 16T @INT8
媒体	16 Channel Decoder - H.264/265 1080P30 1 Channel Encoder
内存	LPDDR4X 2*64bit @3733MT/S
接口	PCIe3.0 / USB 3.0/GE
封装	15mm*15mm
功耗	8TOPS@4W, 16TOPS@8W
工艺	12nm FFC

【应用案例】

华为针对其昇腾 AI 芯片的计算架构专门构建了完整的软件栈，兼容各个深度学习框架并能够高效运行在昇腾 AI 芯片上，让开发者能够快速开发推理应用，为开发者提供便利的解决方案。当前主流的深度学习应用，包括图像分类、人脸识别、目

【产品特征 / 技术特点】

达芬奇架构主要由计算单元、存储系统和控制单元三部分构成。其中计算单元又分为：矩阵计算单元、向量计算单元、标量计算单元，分别对应矩阵、向量和标量三种常见的计算模式。

■ 矩阵计算单元（Cube Unit）：

矩阵计算单元和累加器主要完成矩阵相关运算。一拍完成一个 FP16 的 16x16 与 16x16 矩阵乘（4096）；如果是 INT8 输入，则一拍完成 16*32 与 32*16 矩阵乘（8192）；

■ 向量计算单元（Vector Unit）：

实现向量和标量，或双向量之间的计算，功能覆盖各种基本的计算类型和许多定制的计算类型，主要包括 FP16/FP32/INT32/INT8 等数据类型的计算；

■ 标量计算单元（Scalar Unit）：

相当于一个微型 CPU，控制整个 AI Core 的 以及基本的算术运算。
运行，完成整个程序的循环控制、分支判断，可以 **【专利状态】**（未提供）
为 Cube/Vector 提供数据地址和相关参数的计算，

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
昇腾 310AI 处理器	8TOPS@ 4W16TOPS@8W	LPDDR4X	2*64bit @3733MT/S	2019 年	16T	8T		12nm FFC

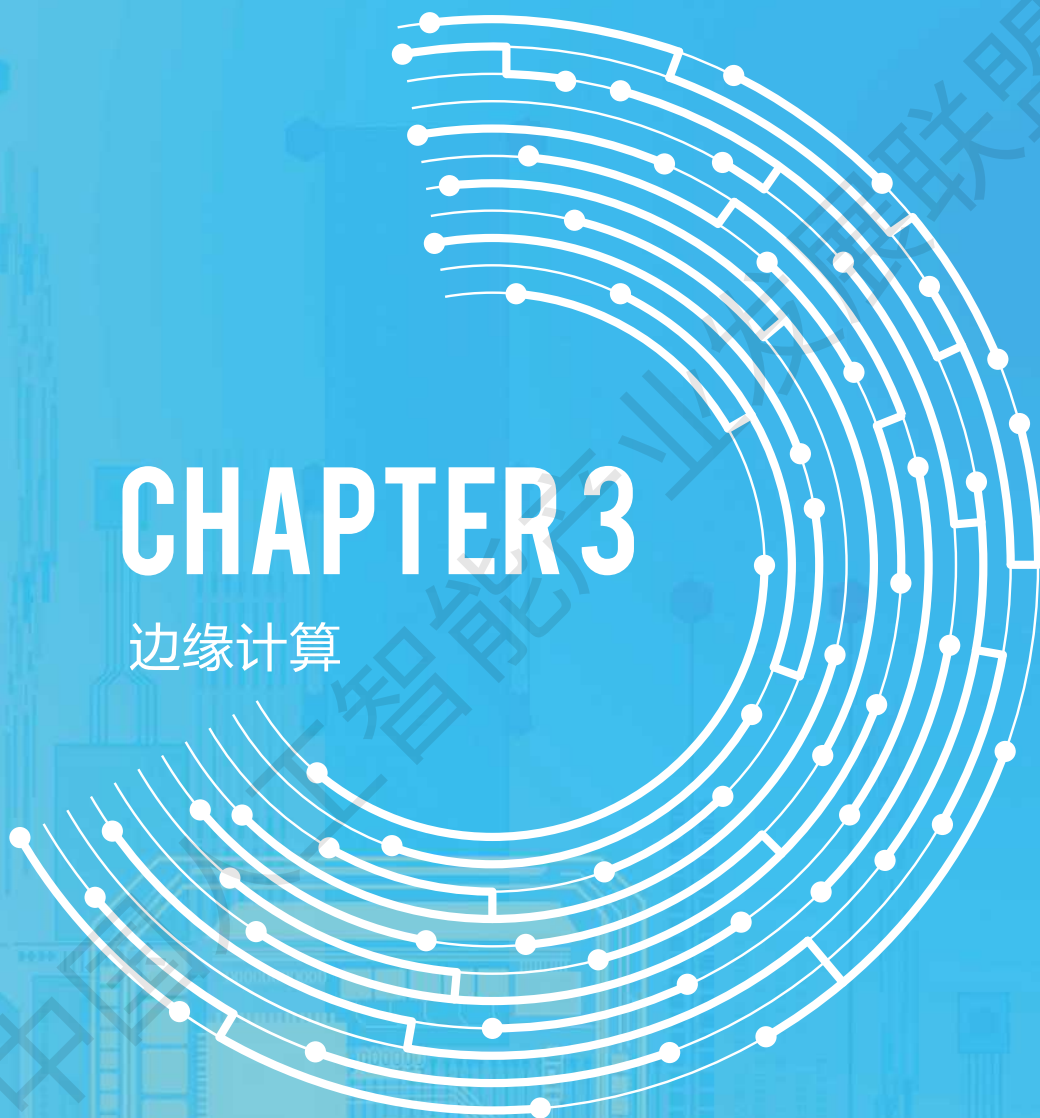
2.3 小结

		公司名称	芯片名称	功耗	内存	带宽
云端 / 数据中心	云端 训练	北京百度网讯科技有限公司	百度昆仑 1	150W	16GB	512GB/s
		上海燧原科技有限公司	邃思	190W (最大功耗)	16GB HBM	512GB /s
		NVIDIA	NVIDIA A100	400W	40GB HBM2	1555 GB/s
			NVIDIA V100(NVLink)	300W	32/16 GB HBM2	900 GB/s
			NVIDIA V100(PCIe)	250	32/16 GB HBM2	900 GB/s
			NVIDIA V100S	250W	32 GB HBM2	1134 GB/s
	云端 推断	NVIDIA	NVIDIA T4	70W	16 GB GDDR6	320+ GB/s
		赛灵思	Alveo U50 (VU35P)	50W	8GB	316GB/s
			Alveo U200 (VU9P)	100W	64GB	77GB/s
			Alveo U250 (VU13P)	110W	64GB	77GB/s
			Alveo U280 (VU37P)	100W	/	/
		中科寒武纪科技股份有限公司	思元 100	75W	8GB/16GB	102GB/s
			思元 270	70W	16GB	102GB/s
		北京比特大陆科技有限公司	BM1684	16W	12GB LPDDR4X	68.3GB/s
		深圳鲲云信息科技有限公司	星空加速卡 X3	峰值 56W	两组 DDR4 SDRAM, 8GB/DDR	32GB/s
		华为技术有限公司	昇腾 310AI 处理器	8TOPS@ 4W16TOPS@8W	LPDDR4X	2*64bit @3733MT/S

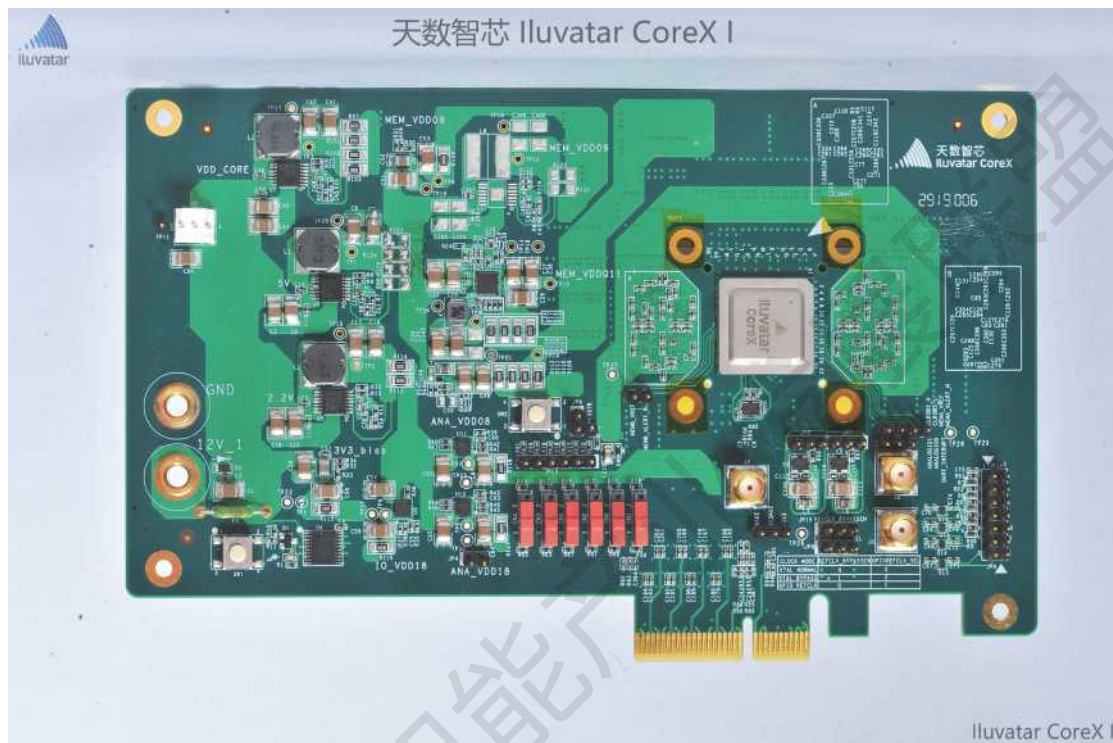
上市时间	深度学习计算能力						工艺
	INT8	FP16	FP32	FP64	TF32	FP64 Tensor Core	
2019 年 11 月	256TOPS	64 TOPS	16 TOPS				14nm
2020 年 6 月	80TOPS	80 TFLOPS	20 TFLOPS				GF 12nm FinFET
2020 年 5 月	"624 TOPS 1248 TOPS* (* 采用稀疏技术)"	312 TF 624 TF* (* 采用稀疏 技术)	19.5 TFOPS	9.7 TFOPS	"312 TFOPS (* 采 用稀疏技术)"	19.5 TFOPS	7nm
2017 年 5 月		125 TFLOPS	15.7 TFLOPS	7.8 TFLOPS			12nm
2017 年 5 月		112 TFLOPS	14 TFLOPS	7 TFLOPS			12nm
2019 年 11 月		130 TFLOPS	16.4TFLOPS	8.2 TFLOPS			12nm
2018 年 3 月	130TOPS	65TFLOPS	8.1TFLOPS				12nm
2019 年 8 月	16.2TOPS	/	/				16nm UltraScale™
2018 年 10 月	18.6TOPS	/	/				16nm UltraScale™
2018 年 10 月	33.3TOPS	/	/				16nm UltraScale™
2018 年 11 月	24.5TOPS	/	/				16nm UltraScale™
2018 年 5 月	32TOPS	16 TFLOPS	/				16nm
2019 年 6 月	128 TOPS	/	/				16nm
2019 年 9 月	"17.6TOPS (普通模式); 35.2TOPS (Winograd 加速模式)"		2.2T				12nm
2020 年 6 月	10.9 TOPS	N/A	N/A				28nm
2019 年	16T	8T					12nm FFC

CHAPTER 3

边缘计算



3.1 上海天数智芯半导体有限公司 – 天数智芯 Iluvatar CoreX I 芯片



【技术概述】

天数智芯 Iluvatar CoreX I 芯片是聚焦于边缘应用的高精度深度学习推理人工智能芯片，针对市场主流 CNN 网络模型算法进行了优化，可以减少存储单元延迟，增加运算并行度，最终提高边缘端人工智能应用的实际性能。

【技术指标】

芯片高效支持 CNN 等卷积类矩阵运算，可应用于图像识别与部分语音识别人工智能应用场景。在 5W 的峰值功耗条件下，可提供 4.8TFLOPS 的半精度浮点 AI 算力 (FP16)，16GB/s 的双向 IO 带宽，与 8GB 的本地存储容量，并已适配包括 ARM、X86 与龙芯等国产 CPU。

【应用案例】

智慧新零售、基于人脸识别的智慧园区、智能垃圾分类、智慧医疗等基于视频与图像识别的应用场景。

【产品特征 / 技术特点】

Iluvatar CoreX I 芯片是基于数据流的深度神经网络处理器架构的协处理器，支持基于 CNN 网络模型的边缘侧人工智能推理。可用于智能制造、智慧新零售、智慧医疗、智慧园区、智能垃圾分类等多个行业及领域的应用，助力产业智能化升级。

该芯片具有以下主要特征：

1. 支持基于深度学习算法的视频图像类推理应用；



2. 主流深度神经网络架构支持，如 ResNet, VGG, YoloV2, SSD, SPP-Net 等；

3. 单芯片 4.8TFLOPS FP16 运算能力，32 路并行计算应对数据处理；

4. 基于本地存储预取的图像推理分析低延迟；

5. 峰值功耗低至 5W 并支持低成本无风扇系统；

6. 接口符合 PCIe4.0 工业标准，易于扩展，可支持芯片间及卡间互联扩展；

7. Tensorflow 原生框架扩展，支持客户开发环境无痛迁移。

【专利状态】

申请 17 件。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Iluvatar CoreX I	峰值 5W	8GB	273Gb/s	2019 年 10 月	NA	4.8T	NA	T16

3.2 NVIDIA – NVIDIA Jetson , NVIDIA T4 Tensor Core GPUs



【技术概述】

NVIDIA Jetson:

NVIDIA Jetson 系统所提供的性能和能效可提高自主机器软件的运行速度，而且功耗更低。每个系统都是一个完备的模块化系统 (SOM)，具备 CPU、GPU、PMIC、DRAM 和闪存。NVIDIA Jetson 具备可扩展性。选择适合应用场合的 SOM，构建自定义系统，满足特定的应用需求。

NVIDIA T4 Tensor Core GPUs:

数据中心 GPU，可加快 AI、高性能计算 (HPC) 和图形技术的发展。

【技术指标】

NVIDIA Jetson:

Jetson Nano 采用 NVIDIA Maxwell 架构，配备 128 个 NVIDIA CUDA 核心。

Jetson TX2 采用 NVIDIA Pascal 架构，配

备 256 个 NVIDIA CUDA 核心。

Jetson Xavier NX 采用 NVIDIA Volta 架构，配备 384 个 NVIDIA CUDA 核心和 48 个 Tensor 核心。

Jetson AGX Xavier 采用 NVIDIA Volta 架构，配备 512 个 NVIDIA CUDA 核心和 64 个 Tensor 核心。

NVIDIA T4 GPU:

NVIDIA Turing GPU 架构成功将实时光线追踪、AI、模拟和光栅化技术集成于一身。

【应用案例】

NVIDIA Jetson:

NVIDIA Jetson 平台是 NVIDIA AI 边缘平台，可以覆盖不同领域、不同行业。目前比较集中应用于两个领域：自主机器和高清传感器及视频分析。其中包括小型商用机器人、无人机、智能高分辨率

传感器（用于工厂物流和生产线）、光学检测、网络录像机，便携式医疗设备以及其他工业物联网（IoT）系统。

NVIDIA T4 GPU:

利用 NVIDIA T4 GPU 处理高性能计算 (HPC) 和超大规模数据中心工作负载。其宣称数据科学家和研究人员可以在能源勘探和深度学习等应用场合中解析 PB 级的数据，速度比使用传统 CPU 快几个数量级。NVIDIA T4 加速器可以超快速度运行更大的模拟。此外，NVIDIA GPU 还能为虚拟桌面、应用程序和 workstation 提供超高性能和用户密度。

【产品特征 / 技术特点】

NVIDIA Jetson:

Jetson Nano 助力开发数百万个新的小型、低功耗的 AI 系统。在嵌入式物联网应用程序领域，包括入门级网络硬盘录像机 (NVR)、家用机器人以

及具备全面分析功能的智能网关等场景得以应用。

Jetson TX2 为用户的嵌入式 AI 计算设备提供丰富的标准硬件接口，适配各类产品和外形规格。

Jetson Xavier NX 是一种外形小巧的模块，它能并行运行多个现代神经网络，处理来自多个高分辨率传感器的数据，进而满足完整 AI 系统的需求。

Jetson AGX Xavier 是专为自主机器设计的计算机。该计算机具有六个计算机主板，用于处理传感器数据和运行自主机器软件，并且可以提供完全自主机器所需的性能和能效。

NVIDIA T4 GPU:

NVIDIA T4 GPU 推理加速器搭载 NVIDIA Turing Tensor Core，提供多精度推理性能，以加速现代人工智能的各种应用。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽		深度学习计算能力			工艺
					INT8	FP16	FP32	
Jetson Nano	5W 10W	4 GB 64-bit LPDDR4	25.6 GB/s	2019 年 3 月		0.5 TFLOPS		
Jetson TX2	7.5W 15W	8 GB 128-bit LPDDR4	58.4 GB/s	2017 年 3 月		1.3 TFLOPS		
Jetson Xavier NX	10W 15W	8 GB 128-bit LPDDR4	51.2 GB/s	2019 年 11 月	14 21 TOPS			
Jetson AGX Xavier	10W 15W 30W	16 GB 256-bit LPDDR4	137 GB/s	2018 年 12 月	32 TOPS	16 TFLOPS		
Jetson AGX Xavier	10W 20W	8 GB 256-bit LPDDR4	85 GB/s	2019 年	19.3 TOPS	9.6 TFLOPS		
NVIDIA T4	70w	16 GB GDDR6	320+ GB/s	2018 年 3 月	130 TOPS	65 TFLOPS	8.1 TFLOPS	12nm

3.3 紫光展锐 – 虎贲 T710

【技术概述】

虎贲 T710 采用 8 核 CPU 架构，由 4 颗 2.0GHz 的 Arm Cortex-A75 及 4 颗 1.8GHz 的 Arm Cortex-A55 组成，搭载工作频率为 800MHz 的 IMG PowerVR GM 9446 图形处理器。虎贲 T710 包含了 CPU、GPU、NPU、ISP、VDSP 等处理单元，为各类丰富的 AI 应用提供了高效能、低功耗的技术基础。

【技术指标】

CPU: 4xA75@2.0GHz + 4xA55@1.8GHz
 GPU: PowerVR GM 9446@800MHz
 LCD: 支持 QHD+ 分辨率 (3200x1440)
 Camera: 支持 24M+8M/16M+16M 双摄
 NPU: 3.2TOPS@INT8 主频 800MHz
 网络制式: GSM/WCDMA/TDSCDMA/
 TDD-LTE/FDD-LTE/
 Wi-Fi: 802.11a,b,g,n,11ac
 Bluetooth: 5.0

【应用案例】

虎贲 T710 已经在智能医疗，智能零售和智能



安防多个领域内得到应用。

【产品特征 / 技术特点】

采用异构双核架构，虎贲 T710 的计算能效比超过 2.5TOPS/W，并支持 INT4, INT8, INT16, FP16 等多种神经网络量化方式。同时还支持 4K@30 帧视频编解码, WiFi 11AC 和 BT 5.0 等多种无线连接技术。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps	2019 年 8 月	3.2TOPS	0.5TOPS		12nm



3.4 中科寒武纪科技股份有限公司 - 寒武纪思元 220 M.2



【技术概述】

基于寒武纪最新一代 MLUv02 架构，实现终端设备和边缘端设备的 AI 赋能。

无人机等边缘计算场景，支持视觉、语音、自然语言处理以及传统机器学习等多样化的人工智能应用，实现各种业务的边缘端智能化解决方案。

【技术指标】

专门用于深度学习的 SOC 芯片，具有高算力、低功耗、丰富 I/O 接口。

【产品特征 / 技术特点】

1. MLUv02 架构，提升缓存有效容量和带宽；
2. U 盘大小尺寸可提供 8 路高清视频分析；
3. 计算可编程，满足多样化 AI 定制要求。

【应用案例】

可应用于智能制造、无人零售、智能交通、

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
思元 220 M.2	8.25W	4GB	29GB/s	2019 年 11 月	8TOPS	/	/	16nm

3.5 华为技术有限公司 – 昇腾 310AI 处理器



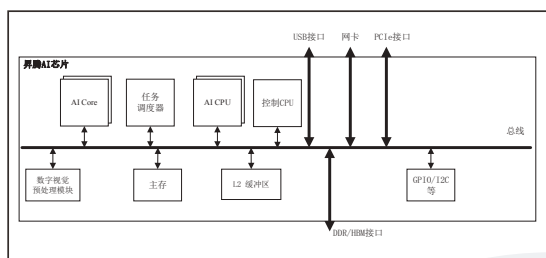
【技术概述】

昇腾 AI 处理器的主要架构组成:

- 芯片系统控制 CPU (Control CPU)
- AI 计算引擎 (包括 AI Core 和 AI CPU)
- 多层级的片上系统缓存 (Cache) 或缓冲区 (Buffer)

区 (Buffer)

- 数字视觉预处理模块 (Digital Vision Pre-Processing, DVPP) 等



流的深度学习应用, 包括图像分类、人脸识别、目标检测、光学字符识别、视频处理和自然语言处理领域的各个模型, 均可以在昇腾 310 处理器上得到很好的技术支持。

【产品特征 / 技术特点】

达芬奇架构主要由计算单元、存储系统和控制单元三部分构成。其中计算单元又分为: 矩阵计算单元、向量计算单元、标量计算单元, 分别对应矩阵、向量和标量三种常见的计算模式。

■ 矩阵计算单元 (Cube Unit) :

矩阵计算单元和累加器主要完成矩阵相关运算。一拍完成一个 FP16 的 16x16 与 16x16 矩阵乘 (4096); 如果是 INT8 输入, 则一拍完成 16*32 与 32*16 矩阵乘 (8192);

■ 向量计算单元 (Vector Unit) :

实现向量和标量, 或双向量之间的计算, 功能覆盖各种基本的计算类型和许多定制的计算类型, 主要包括 FP16/FP32/INT32/INT8 等数据类型的计算;

■ 标量计算单元 (Scalar Unit) :

相当于一个微型 CPU, 控制整个 AI Core 的运行, 完成整个程序的循环控制、分支判断, 可以为 Cube/Vector 提供数据地址和相关参数的计算, 以及基本的算术运算。

【技术指标】

规格	描述
架构	达芬奇
性能	Up to 8T @FP16 Up to 16T@INT8
媒体	16 Channel Decoder - H.264/265 1080P30 1 Channel Encoder
内存	LPDDR4X 2*64bit @3733MT/S
接口	PCIe3.0 /USB 3.0/GE
封装	15mm*15mm
功耗	8TOPS@4W, 16TOPS@8W
工艺	12nm FFC

【应用案例】

华为针对其昇腾 AI 芯片的计算架构专门构建了完整的软件栈, 兼容各个深度学习框架并能够高效运行在昇腾 AI 芯片上, 让开发者能够快速开发推理应用, 为开发者提供便利的解决方案。当前主



【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
昇腾 310AI 处理器	8TOPS@ 4W16TOPS@8W	LPDDR4X	2*64bit @3733MT/S	2019 年	16T	8T		12nm FFC

3.6 深圳鲲云信息科技有限公司 — 星空 x3 加速卡



【技术概述】

星空 X3 加速卡为鲲云推出的面向边缘端和数据中心进行深度学习推断的 AI 计算加速卡，搭载鲲云自研的定制数据流 CAISA 芯片，采用无指令集的架构方式，为支持深度学习的边缘和数据中心服务器提供计算加速方案。

【技术指标】

1. 支持 ResNet、VGG、YOLO 等多个主流 CNN 网络
2. 提供 RainBuilder 编译工具链，支持端到端算法开发和部署
3. 支持 TensorFlow、Caffe、PyTorch 及 ONNX (MXNet) 等主流深度学习框架开发的算法模型
4. 其宣称芯片利用率可达 95.4%

【应用案例】

1. 工业领域：鲲云与合作方针对某电容上激光刻蚀字符推出基于深度学习的 OCR 识别方案，识别算法采用了深度学习方法，可识别字母、数字以及数十种特殊字符，字符识别精度可达到 99.99%。采用搭载鲲云 CAISA 芯片的星空加速卡，识别延迟可以降低到 50ms，识别速度可达到 20pcs/s。

2. 电力领域：

鲲云与合作方通过研究适用于电力无人机巡检的目标检测深度学习算法，实现基于人工智能技术的电力无人机智能巡检、数据采集，并在服务器端进行高质量的图片数据分析，提高巡检效率，降低巡检工作量。

【产品特征 / 技术特点】

- 1 其宣称芯片利用率可达 95.4%；

2 时延：3ms 分类延时；

3 支持分类、目标检测以及语义分割类深度学习算法；

4 Batch size 不敏感；

5 温度范围：-20℃ ~70℃

6 实测 Benchmark:

算法网络	ResNet-50	ResNet-152	YOLOv3	SSD-ResNet50	U-Net Industrial
延时	3.05 ms	8.68 ms	31.06 ms	20.8 ms	74.07 ms
吞吐率	1306.93 FPS	460.27 FPS	125.75 FPS	182.16 FPS	54.01 FPS
芯片利用率	92.32%	95.43%	82.37%	77.06%	64.97%

注：Batch=4，INT 8

【专利状态】

已申请专利 110 件，包括 109 件发明专利、1 件外观设计专利；

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
星空加速卡 X3	峰值 56W	两组 DDR4 SDRAM, 8GB	32Gbps	2020 年 6 月	10.9 TOPS	N/A	N/A	28nm

中国人工智能产业发展联盟

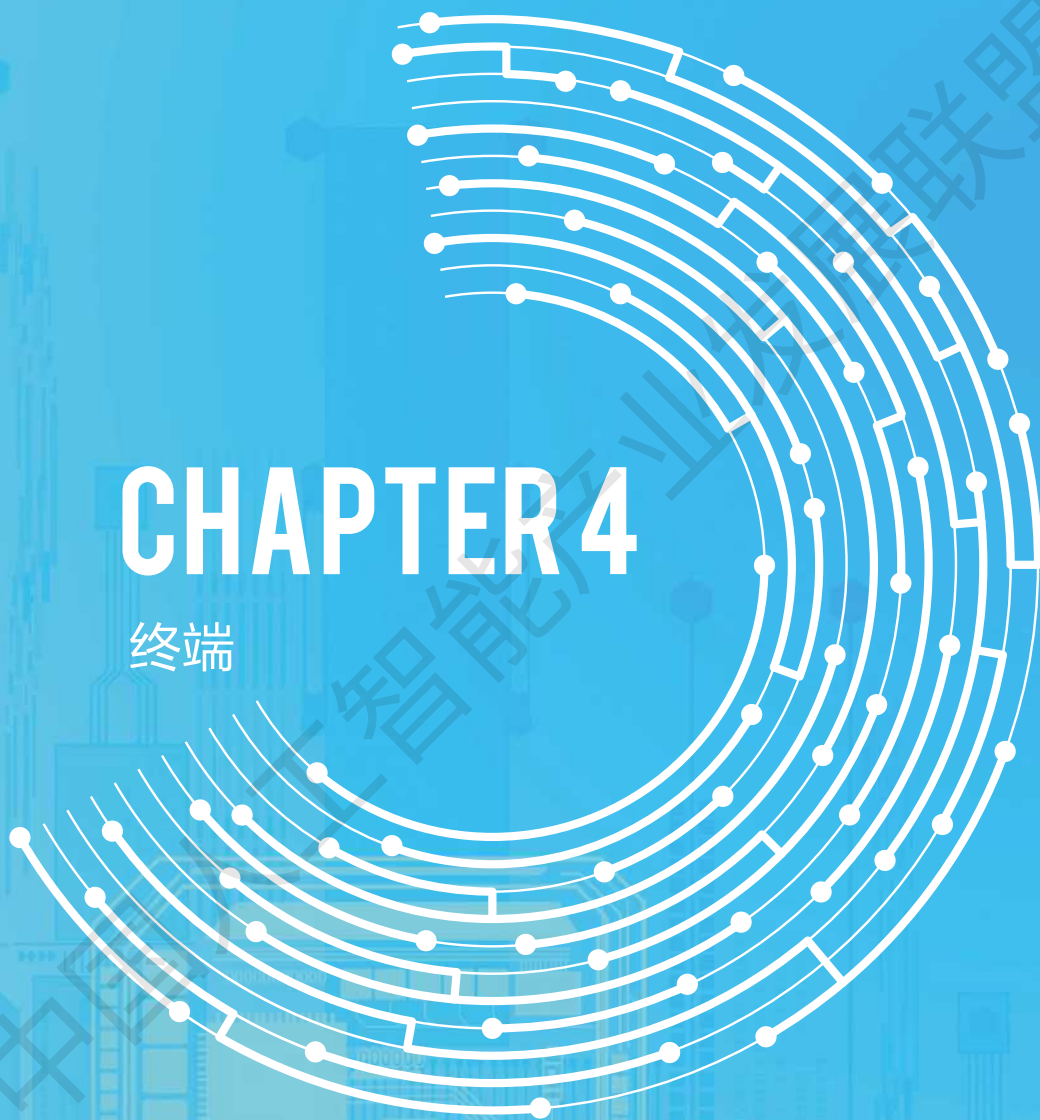
3.7 小结

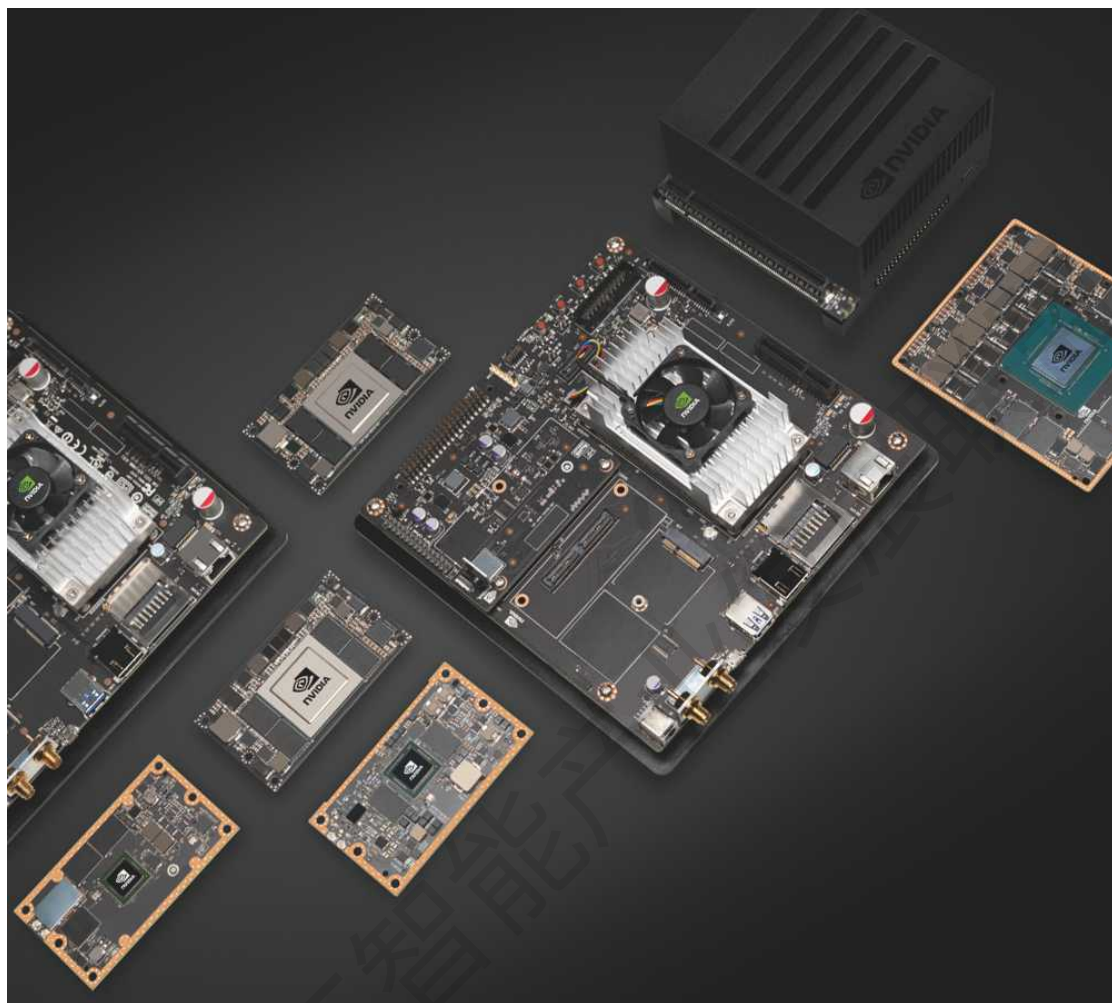
边缘 计算	公司名称	芯片名称	功耗	内存	带宽
	上海天数智芯 科技有限公司	Iluvatar CoreX I	峰值 5W	8 GB	273Gb/s
	NVIDIA	Jetson Nano	5W 10W	4 GB 64-bit LPDDR4	25.6GB/s
		Jetson TX2	7.5W 15W	8 GB 128-bit LPDDR4	58.4GB/s
		Jetson Xavier NX	10W 15W	8 GB 128-bit LPDDR4	51.2 GB/s
		Jetson AGX Xavier	10W 15W 30W	16 GB 256-bit LPDDR4	137GB/s
		Jetson AGX Xavier	10W 20W	8 GB 256-bit LPDDR4	85GB/s
		NVIDIA T4	70W	16 GB GDDR6	320+ GB/s
	紫光展锐	虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps
	中科寒武纪科技 股份有限公司	思元 220 M.2	8.25W	4GB	29GB/s
	华为技术有限公司	昇腾 310AI 处理 器	"8Tops@4W 16Tops@8W"	LPDDR4X	2*64bit @3733MT/S
	深圳鲲云信息 科技有限公司	星空加速卡 X3	峰值 56W	两 组 DDR4 SDRAM, 8GB/DDR	32GB/s

上市时间	深度学习计算能力						工艺
	INT8	FP16	FP32	FP64	TF32	FP64 Tensor Core	
2019 年 10 月	NA	4.8TFLOPS	NA				T16
2019 年 3 月		0.5 TFLOPS					
2017 年 3 月		1.3TFLOPS					
2019 年 11 月	14 21TOPS						
2018 年 12 月	32TOPS	16TFLOPS					
2019 年	19.3TOPS	9.6TFLOPS					
2018 年 3 月	130TOPS	65TFLOPS	8.1TFLOPS				12nm
2019 年 8 月	3.2TOPS	0.5TOPS					12nm
2019 年 11 月	8TOPS	/	/				16nm
2019 年	16TOPS	8TFLOPS					12nm FFC
2020 年 6 月	10.9 TOPS	N/A	N/A				28nm

CHAPTER 4

终端





4.1 通用端侧产品

4.1.1 NVIDIA – NVIDIA Jetson



【技术概述】

NVIDIA Jetson 系统所提供的性能和能效可提高自主机器软件的运行速度，而且功耗更低。每个系统都是一个完备的模块化系统 (SOM)，具备 CPU、GPU、PMIC、DRAM 和闪存。NVIDIA Jetson 具备可扩展性。通过选择适合应用场合的 SOM，即能够以此为基础构建自定义系统，满足特定的应用需求。

【技术指标】

Jetson Nano 采用 NVIDIA Maxwell 架构，配备 128 个 NVIDIA CUDA 核心。

Jetson TX2 采用 NVIDIA Pascal 架构，配备 256 个 NVIDIA CUDA 核心。

Jetson Xavier NX 采用 NVIDIA Volta 架构，配备 384 个 NVIDIA CUDA 核心和 48 个 Tensor 核心。

Jetson AGX Xavier 采用 NVIDIA Volta 架构，配备 512 个 NVIDIA CUDA 核心和 64 个 Tensor 核心。

【应用案例】

NVIDIA Jetson 平台是 NVIDIA AI 边缘平台，可以覆盖不同领域、不同行业。目前比较集中应用于两个领域：自主机器和高清传感器及视频分析。其中包括小型商用机器人、无人机、智能高分辨率传感器（用于工厂物流和生产线）、光学检测、网络录像机，便携式医疗设备以及其他工业物联网（IoT）系统。

【产品特征 / 技术特点】

Jetson Nano 助力开发数百万个新的小型、

低功耗的 AI 系统。在嵌入式物联网应用程序领域，包括入门级网络硬盘录像机（NVR）、家用机器人以及具备全面分析功能的智能网关等场景得以应用。

Jetson TX2 为用户的嵌入式 AI 计算设备提供相应的速度和能效。提供丰富的标准硬件接口，适配各类产品和外形规格，实现的 AI 计算终端。

Jetson Xavier NX 外形小巧，它能并行运行多个现代神经网络，处理来自多个高分辨率传感器的数据，进而满足 AI 系统的需求。

Jetson AGX Xavier 是专为自主机器设计的计算机。该计算机具有六个计算机主板，用于处理传感器数据和运行自主机器软件，并且可以提供完全自主机器所需的性能和能效。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Jetson Nano	5W 10W	4 GB 64-bit LPDDR4	25.6GB/s	2019 年 3 月		0.5 TFLOPS		
Jetson TX2	7.5W 15W	8 GB 128-bit LPDDR4	58.4GB/s	2017 年 3 月		1.3TFLOPS		
Jetson Xavier NX	10W 15W	8 GB 128-bit LPDDR4	51.2 GB/s	2019 年 11 月	14 21 TOPS			
Jetson AGX Xavier	10W 15W 30W	16GB 256-bit LPDDR4	137 GB/s	2018 年 12 月	32 TOPS	16TFLOPS		
Jetson AGX Xavier	10W 20W	8GB 256-bit LPDDR4	85GB/s	2019 年	19.3 TOPS	9.6TFLOPS		

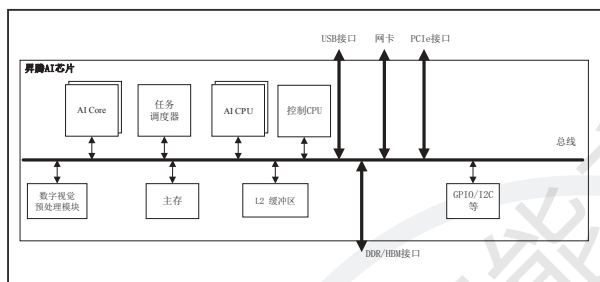
4.1.2 华为技术有限公司 – 昇腾 310AI 处理器



【技术概述】

昇腾 AI 处理器的主要架构组成：

- 芯片系统控制 CPU（Control CPU）
- AI 计算引擎（包括 AI Core 和 AI CPU）
- 多层级的片上系统缓存（Cache）或缓冲区（Buffer）
- 数字视觉预处理模块（Digital Vision Pre-Processing, DVPP）等



在昇腾 AI 芯片上，让开发者能够快速开发推理应用，为开发者提供便利的解决方案。当前主流的深度学习应用，包括图像分类、人脸识别、目标检测、光学字符识别、视频处理和自然语言处理领域的各个模型，均可以在昇腾 310 处理器上得到很好的技术支持。

【产品特征 / 技术特点】

达芬奇架构主要由计算单元、存储系统和控制单元三部分构成。其中计算单元又分为：矩阵计算单元、向量计算单元、标量计算单元，分别对应矩阵、向量和标量三种常见的计算模式。

■ 矩阵计算单元（Cube Unit）：

矩阵计算单元和累加器主要完成矩阵相关运算。一拍完成一个 FP16 的 16x16 与 16x16 矩阵乘（4096）；如果是 INT8 输入，则一拍完成 16x32 与 32x16 矩阵乘（8192）；

■ 向量计算单元（Vector Unit）：

实现向量和标量，或双向量之间的计算，功能覆盖各种基本的计算类型和许多定制的计算类型，主要包括 FP16/FP32/INT32/INT8 等数据类型的计算；

【技术指标】

规格	描述
架构	达芬奇
性能	Up to 8T @FP16 Up to 16T@INT8
媒体	16 Channel Decoder – H.264/265 1080P30 1 Channel Encoder
内存	LPDDR4X 2*64bit @3733MT/S
接口	PCIe3.0 /USB 3.0/GE
封装	15mm*15mm
功耗	8 TOPS @ 4 W , 16TOPS@8W
工艺	12nm FFC

【应用案例】

华为昇腾 AI 芯片的计算架构具有完整的软件栈，其宣称兼容各个深度学习框架并能够高效运行



■ 标量计算单元 (Scalar Unit) :

相当于一个微型 CPU，控制整个 AI Core 的运行，完成整个程序的循环控制、分支判断，可以

为 Cube/Vector 提供数据地址和相关参数的计算，以及基本的算术运算。

【专利状态】 (未提供)

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
昇腾 310AI 处理器	8TOPS @4W16TOPS @8W	LPDDR4X	2*64bit @3733MT/S	2019 年	16T	8T		12nm FFC

4.1.3 瑞芯微电子股份有限公司 – RK1808 AIoT 芯片方案



【技术概述】

RK1808 芯片采用 22 纳米 FDX 工艺，内嵌高性能的 64 位双核 Cortex-A35 CPU 和在人工智能应用场景下深度学习大量使用的神经网络运算加速引擎（NPU），实现运算力的同时满足端侧产品的功耗需求，达到功耗、成本、性能的有效平衡。其宣称具有高性能的神经网络运算单元的设计；高性能丰富接口的复杂 SoC 架构设计；多形态灵活可扩展等特性。

【技术指标】

1. 双核 Cortex-A35，最高频率 1.6GHz
2. NPU 支持 INT8/INT16/FP16；NPU 算力：3 TOPS for INT8 / 300 GOPS for INT16 / 100 GFLOPS for FP16；
3. 支持 RKNN/OpenCL/OpenVX 编程接口；
4. 支持 TensorFlow、Caffe、ONNX、Darknet 模型；
5. 内存支持 800MHz 32-bit LPDDR2/LPDDR3/DDR3/DDR3L/DDR4；

6. 支持 Serial SPI NOR/NAND Flash, EMMC；
7. 视频支持 1080p@60FPS H.264 解码，1080p@30FPS H.264 编码；
8. ISP 支持 2MPixel，支持 AE/AWB/AF；
9. 视频输入 4-lane，MIPI-CSI/MIPI-DSI，支持 Virtual Channel，支持 BT.601/BT.656/BT.1120；
10. 视频输出支持 4-lane，MIPI-DSI，最大分辨率 1920*1080；8-bit Parallel RGB panel，最大到 1280*720；
11. 支持支持 USB3.0/PCIe2.1；内置 2-ch & 8-ch I2S & 8-ch PDM，内置 VAD；支持千兆以太网；
12. 支持 8 x UART/3 x SPI/6 x I2C/11 x PWM/4 x SARADC 等丰富低速外设接口；

【应用案例】

1. Toybrick 算力棒，算力卡；
2. 致远电子 M1808 工业 AI 板；
3. 点创 ADAS/DSM 设备；

4. 动作识别智能摄像头；

【产品特征 / 技术特点】

RK1808 芯片支持目标图像识别、视觉处理、语音识别、语义分析等人工智能应用，设计乘加运算 (MAC) 架构及高效的调度单元，能满足目前大部分算法和应用对较高神经网络推理算力和效率的

需求，可以发挥硬件运算单元的能力，在单位时钟周期内尽可能让更多的运算单元处于工作状态，同时兼容考虑各种不同算法或模型的情况。该项目设计的人工智能硬件加速引擎，可以达到理论上单芯片 3.0TOPS 的运算能力。

【专利状态】（未提供）

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	INT16	FP16	
RK1808	3W	LPDDR2/LPDDR3 DDR3/DDR3L/DDR4	6.4 GBytes/s	2019 年 6 月	3 TOPS	300 GOPS	100 GFLOPS	22nm

4.1.4 瑞芯微电子股份有限公司 – RK3399 Pro 人工智能端侧芯片方案

【技术概述】

RK3399 Pro CPU 采用 big.LITTLE 大小核架构，双核 Cortex-A72+ 四核 Cortex-A53，在整体性能、功耗方面具有技术领先性，GPU 采用四核 ARM 新一代高端图像处理器 Mali-T860，集成更多带宽压缩技术，整体性能优异，适合于电子白板、电子书包，人脸识别设备、无人机、机器人、游戏终端、游戏外设类、手游挂机服务器、家电类、广告机/一体机、金融 POS 类、车载控制业、瘦客户机（云端服务）、VOIP 视频会议系统、教育类平板、卡拉 OK 娱乐类、医疗类、安防/监控/警务、工控类、IoT 物联网领域、VR 录像、VR 等近百行业应用产品。

【技术指标】

1. 双 Cortex-A72+ 四 Cortex-A53 大小核 CPU 结构，频率最高 1.8GHz；
2. GPU 为 四核 Mali-T860，支持 OpenGL ES1.1/2.0/3.0/3.1，OpenVG1.1，OpenCL，DX11；支持 AFBC（帧缓冲压缩）；
3. NPU 支持 INT8/INT16/FP16；NPU 算力：3 TOPS for INT8 / 300 GOPS for INT16 / 100 GFLOPS for FP16；
4. 内存支持双通道 DDR3-1866/DDR3L-1866/LPDDR3-1866/LPDDR4；支持 eMMC 5.1，SDIO3.0；
5. 支持 4K VP9 and 4K 10bits H265/H264 视频解码，60fps；
6. 1080P 多格式视频解码（VC-1，MPEG-1/2/4，VP8）；

7. 1080P 视频编码，支持 H.264，VP8 格式；
8. 视频后期处理器：反交错、去噪、边缘/细节/色彩优化；
9. 双 VOP 显示：分辨率分别支持 4096x2160 及 2560x1600；支持双通道 MIPI-DSI（每通道 4 线）；
10. 显示支持：eDP 1.3（4 线，10.8Gbps）；HDMI 2.0 支持 4K 60Hz 显示，支持 HDCP 1.4/2.2；支持 DisplayPort 1.2（4 线，最高支持 4K 60Hz）；支持 Rec.2020 和 Rec.709 色域转换；
11. 接口支持双 ISP 像素处理能力 13MPix/s，支持双路摄像头数据同时输入；支持 USB3.0 Type-C 接口；支持 PCIe 2.1（4 full-duplex lanes）；内置低功耗 MCU；支持 8 路数字麦克风阵列输入；

【应用案例】

RK3399Pro 平台联合黑芝麻提供 ADAS/DSM 方案，充分利用 3399Pro 的 AI 推理及多媒体性能，支持 6 路视频的输入和 AI 分析，实现了 ADAS 辅助驾驶及 DSM 驾驶员疲劳检测功能，增强了交通运输车辆的安全系数。

【产品特征 / 技术特点】

RK3399Pro 平台高性能、高扩展、全能型的应用特性，技术特点体现在多显示接口、GPU、音视频解码、人工智能以及与用于多终端互联、人机交互上。

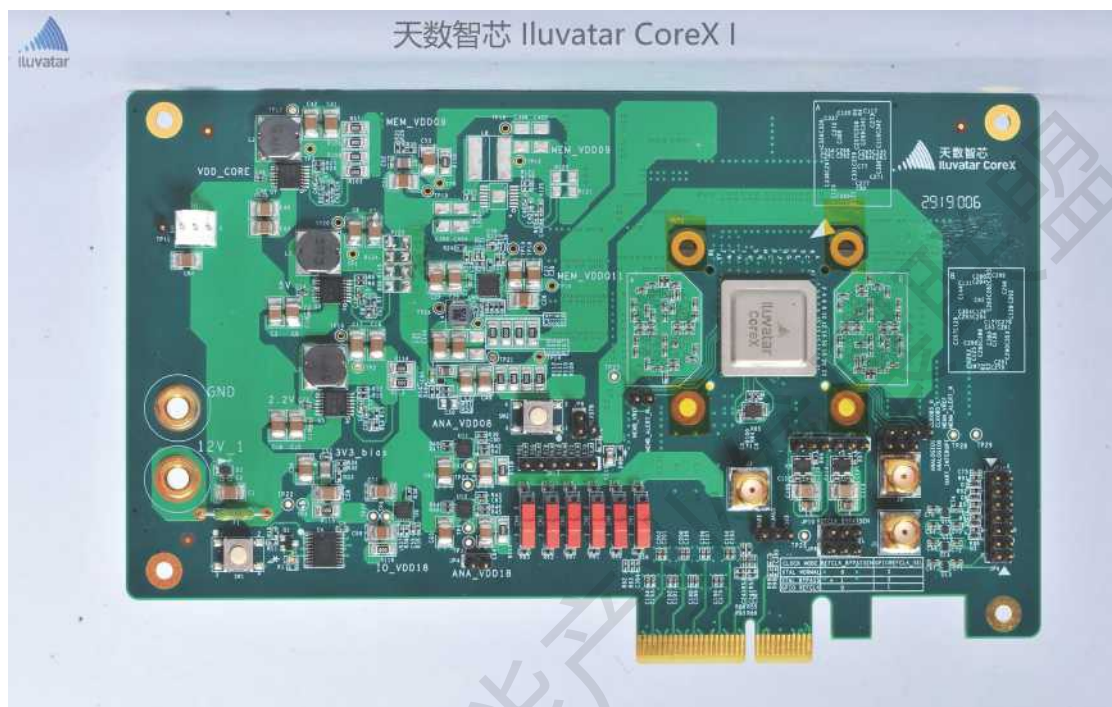
【专利状态】（未提供）



【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	INT16	FP16	
RK3399Pro	10W	双 通 道 DDR3-1866/ DDR3L-1866/ LPDDR3-1866/ LPDDR4	14.9GBytes/s	2018 年 1 月	3T	0.3T	0.1T	28nm

4.1.5 上海天数智芯半导体有限公司 – 天数智芯 Iluvatar CoreX I 芯片



【技术概述】

天数智芯 Iluvatar CoreX I 芯片是聚焦于边缘应用的高精度深度学习推理人工智能芯片，针对市场主流 CNN 网络模型算法进行了优化，可以有效减少存储单元延迟，增加运算并行度，最终大幅提高边缘端人工智能应用的实际性能。

【技术指标】

芯片支持 CNN 等卷积类矩阵运算，可应用于图像识别与部分语音识别人工智能应用场景。在 5W 的峰值功耗条件下，可提供 4.8TFLOPS 的半精度浮点 AI 算力 (FP16)，16GB/s 的双向 IO 带宽，与 8GB 的本地存储容量，并已适配包括 ARM、X86 与龙芯等国产 CPU。

【应用案例】

智慧新零售、基于人脸识别的智慧园区、智能

垃圾分类、智慧医疗等基于视频与图像识别的应用场景。

【产品特征 / 技术特点】

Iluvatar CoreX I 芯片是基于数据流的深度神经网络处理器架构的协处理器，支持基于 CNN 网络模型的边缘侧人工智能推理，具有高精度、高性能、高效能、易扩展、易开发等多项特点。可用于智能制造、智慧新零售、智慧医疗、智慧园区、智能垃圾分类等多个行业及领域的应用，助力产业智能化升级。

该芯片具有以下主要特征：

1. 支持基于深度学习算法的视频图像类推理应用；
2. 主流深度神经网络架构支持，如 ResNet, VGG, YoloV2, SSD, SPP-Net 等；
3. 单芯片 4.8TFLOPS FP16 高精度高性能



运算能力，32 路并行计算应对海量数据处理；

4. 基于本地存储预取的图像推理分析低延迟；

5. 峰值功耗低至 5W 并支持低成本无风扇系统；

6. 接口符合 PCIe4.0 工业标准，易于扩展，可支持芯片间及卡间互联扩展；

7. Tensorflow 原生框架扩展，支持客户开发环境无痛迁移。

【专利状态】

申请 17 件。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Iluvatar CoreX I	峰值 5W	8GB	273Gb/s	2019 年 10 月	NA	4.8T	NA	T16

4.2 智能驾驶

4.2.1 赛灵思

– ZU7EV (Zynq® UltraScale+™ MPSoC)



(ZCU104 板卡



)

基于这两款芯片，赛灵思可提供：从支持边缘传感器的小型器件到用于集中式域控制器的新型高性能器件，通过全方位的产品线，满足汽车对安全、质量和可靠性的各种要求。

【技术指标】

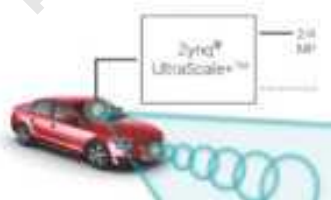
Zynq® UltraScale+™ MPSoC 系列采用 16nm 制造工艺，在单个器件中高度集成 64 位 ARM 处理器和 Xilinx 可编程逻辑 (PL) UltraScale 架构，将实时控制与软硬件引擎相结合，支持图形、视频、波形与数据包处理。Zynq® UltraScale+™ MPSoC 系列更可以广泛用于各种端侧应用，包括智慧城市，辅助驾驶，自动驾驶，医疗临床 / 诊断，工业相机和智慧零售等领域应用。

在 AD/ADAS 场景下，从分布式智能传感器、集中式多传感器融合系统到高度集成的域控制器开发，XA Zynq® UltraScale+™ MPSoC 的并行处理和集成能力既能帮助用户实现高度灵活的解决方案，又能充分满足车辆系统对各种安全及设计需求。

【技术概述】

赛灵思推出的两款车规级芯片 (XA) Zynq® UltraScale+™ MPSoC7EV 和 11EG，可提供优异的可编程能力、性能和 I/O 功能，面向 L2+ 到 L4 的高级驾驶员辅助系统 (ADAS) 应用和自动驾驶 (AD) 应用提供高速数据汇总、预处理和分配 (DAPD) 功能以及计算加速功能。

前视摄像头



GEN 3: Zynq® UltraScale+™ MPSoC

摄像头：4 M 像素

保护范围更大：例如对行人和骑车人的保护
汽车便捷性控制：例如塞车小帮手

Xilinx 价值：

- 异构处理器
- 紧密耦合的应用软件及自定义硬件加速器
- FuSa 功能的安全岛

3D 环绕视图



GEN 3: Zynq® UltraScale+™ MPSoC

障碍物探测

显示：全高清及以上

摄像头：4ch+ 2MPixel

– 灵活应变度更高

特性：

- 动态 3D 周围环境视图（飞速摄像头）
- 高分辨率图形动画
- 高级拖车挂接装置
- 传感器融合
- 机器视觉目标检测，可为低速行驶实现车辆控制和自动紧急刹车（AEB）
- 符合 FuSa 标准

结合 VITIS 统一软件平台提供的深度学习处理器，可实现对 AD/ADAS 场景下各类神经网络模型加速。

基于 Zynq® UltraScale+™ MPSoC 的 DPU 性能

Device	Peak TOPS	Typical Power (Watt)	XA (functional safety qualified)	Hardened video codec
ZU2	0.5	5	Y	
ZU3	0.9	6	Y	
ZU4	1.2	7	Y	Y
ZU5	1.4	7.5	Y	Y
ZU6	2.7	10		
ZU7	2.7	10	Y	Y
ZU9	4.1	12		
ZU11	4.3	16	Y	
ZU15	5.4	/		

【应用案例】

赛灵思 – 百度 ACU 自主泊车专用车载计算平台

赛灵思车规级芯片平台 Zynq® UltraScale+™ MPSoC 正在为百度量产型自主泊车（Automated Valet Parking, AVP）专用车载计算平台 ACU-Advanced（Apollo Computing Unit）提供强大动力。

ACU-Advanced 专门针对自主泊车的特定场景和功能而设计。AVP 应用需要先进且强大的深度学习推断能力来处理复杂的驾驶环境，量产型 ACU-Advanced 借助赛灵思 Zynq UltraScale+ EV 平台强大的传感器融合和 AI 处理能力，取代了用于概念验证（proof-of-concepts, POC）的 GPU。而且，该平台与百度深度学习平台框架百度飞桨（Baidu PaddlePaddle）完全兼容，并包括 5 个摄像头和 12 个超声波雷达，支持零下 40 度至零上 85 度的工作温度范围，可以完全满足严苛的车规量产要求。



图 1: 基于赛灵思车规级芯片平台 Zynq® UltraScale+™ MPSoC 的百度量产型自主泊车专用车载计算平台 ACU-Advanced

图 2: 赛灵思车规级芯片平台 Zynq® UltraScale+™ MPSoC

ACU-Advanced 也是业界首款基于赛灵思 Zynq UltraScale+ MPSoC 5EV 器件而量产的 AVP 专用车载计算平台。

赛灵思 – 戴姆勒 合作 MBUX 车载系统

MBUX 车载系统 (interior assistant) 利用人工智能 (AI) 快速准确地识别车主的身份、情绪、手势以及车内偏好等, 让车辆可以预测驾驶员和乘客的需求, 并且无需使用任何按钮。如“探照灯”功能, 用户晚上想看车内情况, MBUX Interior assistant 可以理解用户的这个需求, 用户只需伸手臂就可自动打开或者关闭灯光来响应。MBUX

车载系统还包含一个带触摸屏操作、增强现实技术导航显示、智能语音控制等功能的高分辨率宽屏驾驶舱。

车内驾驶员及乘客监控系统在依靠 AI 推断的同时受功耗限制, 需要节能解决方案, 特别是在受到不良热环境限制时。同时, 系统还需要低时延来提供对使用者手势的快速响应。赛灵思车规级 Zynq® UltraScale+™ MPSoC 是人工智能加速的理想平台, 可提供更高的灵活性, 满足定制车内体验的需求。

MBUX 车内助手功能:

1. 基于人工智能的手势输入系统, 由 Zynq UltraScale+ MPSoC 提供支持
2. 基于数据流的 AI 引擎可通过避免存储数据来减少时延
3. 识别乘客的自然动作, 以便汽车预测司机和乘客的请求
4. 区分司机和乘客的手势
5. 对肢体语言做出反应, 使舒适功能自动化

【产品特征 / 技术特点】

XA Zynq®UltraScale MPSoC 提供了高度可编程容量、性能和 I/O 能力, 实现了高速、数据聚合、预处理和分发 (DAPD), 以及 L2+ 到 L4 高级驾驶员辅助系统 (ADAS) 和自动驾驶 (AD) 应用的计算加速。

XA Zynq®UltraScale MPSoC 产品列表:

可编程逻辑 (PL)

	ZU5EV	ZU7EV	ZU11EG
系统逻辑单元	256	504	653
内存 (Mb)	23.1	38.0	43.6
DSP Slice	1,248	1,728	2,928
视频代码单元 (VCU)	1	1	
最大 I/O 引脚	252	204	464



处理系统（PS）

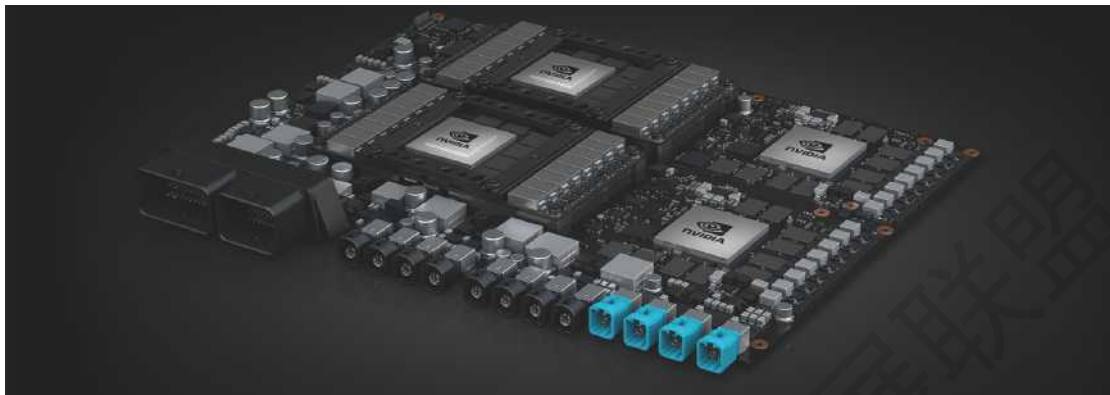
特性	所有器件
应用处理单元	CoreSight 四核 ARM Cortex-A53 MPCore; NEON 和单 / 双高精度浮点; 32KB/32KB L1 Cache、1MB L2 Cache
实时处理单元	CoreSight 双核 ARM Cortex-R5; 单 / 双高精度浮点; 32KB/32KB L1 Cache 和 TCM
嵌入式和外部	256KB 片上内存 w/ECC; 外部 DDR4; DDR3; DDR3L; LPDDR4; LPDDR3; 外部 Quad-SPI; NAND; eMMC
通用连接	214 PS I/O; UART; CAN; USB 2.0; I2C; SPI; 32b GPIO; 实时时钟; WatchDog 计时器; 三态计时器
高速连接功能	4 PS-GTR; PCIe Gen1/2; Serial ATA 3.1; DisplayPort 1.2a; USB 3.0; SGMII
图像处理单元	ARM Mali™-400 MP2; 64KB L2 Cache
应用处理单元	CoreSight 四核 ARM Cortex-A53 MPCore; NEON 和单 / 双高精度浮点; 32KB/32KB L1 Cache、1MB L2 Cache

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
ZU7EV	10	/	/	2019.11	/	/	/	16 nm UltraScale™
ZU11EG	16	/	/	2019.11	/	/	/	

4.2.2 NVIDIA – NVIDIA DRIVE



【技术概述】

DRIVE AGX Pegasus 和 DRIVE AGX Xavier 是 NVIDIA DRIVE 自动驾驶运算硬件平台，由自动驾驶处理芯片 Xavier 构成。Xavier 拥有超过 90 亿个晶体管，可以处理海量数据。Xavier SOC 专为 AI 超级计算而打造。

【技术指标】

自动驾驶处理芯片 Xavier 内置六种处理器：ISP（图像信号处理器）、VPU（视频处理单元）、PVA（可编程视觉加速器）、DLA（深度学习加速器）、CUDA GPU 和 CPU，每秒可进行近 40 万亿次运算，深度学习 30 万亿次。这一处理水平比上一代 DRIVE PX2 参考设计要强大 10 倍，可应用于先进的生产车型中。

【应用案例】

Xavier 系统级芯片架构适用于自动驾驶汽车

【产品信息】

运算平台名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
DRIVE AGX Pegasus	300W	LPDDR4 + GDDR6	1 TB/s	2018 年 11 月	320 TOPS			12nm
DRIVE AGX Xavier	30W	16 GB 256-bit LPDDR4	137 GB/s	2018 年 11 月	30 TOPS			12nm

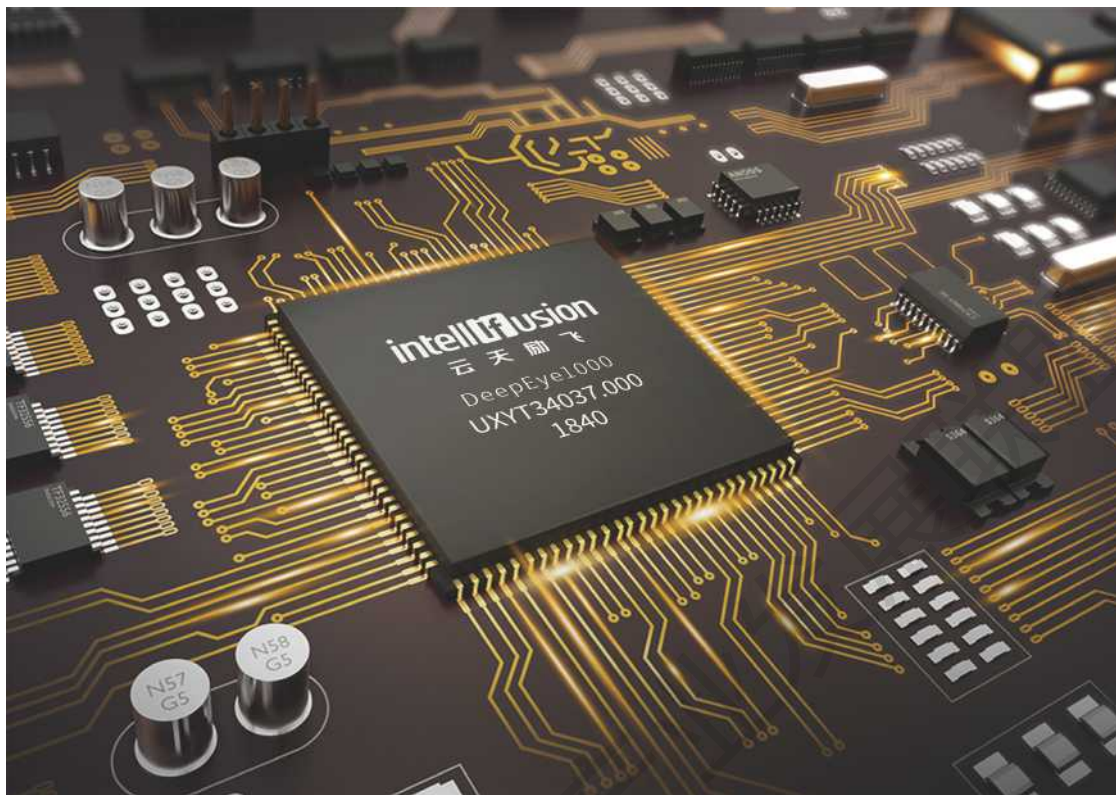
的应用程序，NVIDIA 也一直致力于实现安全的自动驾驶。

【产品特征 / 技术特点】

DRIVE AGX Pegasus 采用基于两个 NVIDIA Xavier 处理器和两个新一代 TensorCore GPU 的架构，实现了超快速的深度学习性能（达到 320 TOPS）。这台高效能、高性能的 AI 计算机同时运行一系列深度神经网络，设计为安全处理高度自动化的驾驶和全自动驾驶。

DRIVE AGX Xavier 提供 30 TOPS 的性能，功耗为 30 瓦。Xavier SoC 在设计上注重安全性，它包含六种不同类型的处理器，可高效运行多种冗余性深度学习算法。

【专利状态】（未提供）



4.3 安防、机器人、车载

4.3.1 深圳云天励飞技术有限公司 – 面向计算机视觉的深度学习神经网络处理器

【技术概述】

（以下简称 DeepEye1000）是一款面向视觉的，自主可编程，用于深度学习神经网络推理加速的，高性能异构 SoC 处理器芯片。

【技术指标】

云天励飞自主产权神经网络处理器（NNP）采用自研 ASIP 深度定制指令集，支持灵活可编程计算流；支持混合精度，可保证算法模型精度；通过计算存储深度融合，配合智能算子引擎、动态任务调度、高效存储调度，可以大幅减少外部存储访问带宽，可以高效执行各种深度算法的推理计算，实现高性能、低功耗的 CNN 网络模型

推理计算加速；

采用阿里平头哥玄铁处理器（CPU）负责整个 SoC 系统的任务调度和分配；视觉数字信号处理器（DSP）负责通用 OpenCV 视觉处理算法执行；视频编码，图像编解码负责完成 H.265/H.264/MJPEG 等多种格式的视频及图片解码功能；

DeepEye1000 支持丰富的外设接口，包括 USB3.0/USB2.0、SDIO、UART、I2C 等；

为了方便 DeepEye1000 的使用，云天励飞为 DeepEye1000 提供了基于 TVM 打造的芯片工具链和软件开发包（SDK），支持 Caffe，MxNET，TensorFlow，ONNX 等深度学习框架，支持一键部署；

【应用案例】

1. DeepEye1000 智能视觉后置协处理解决方案

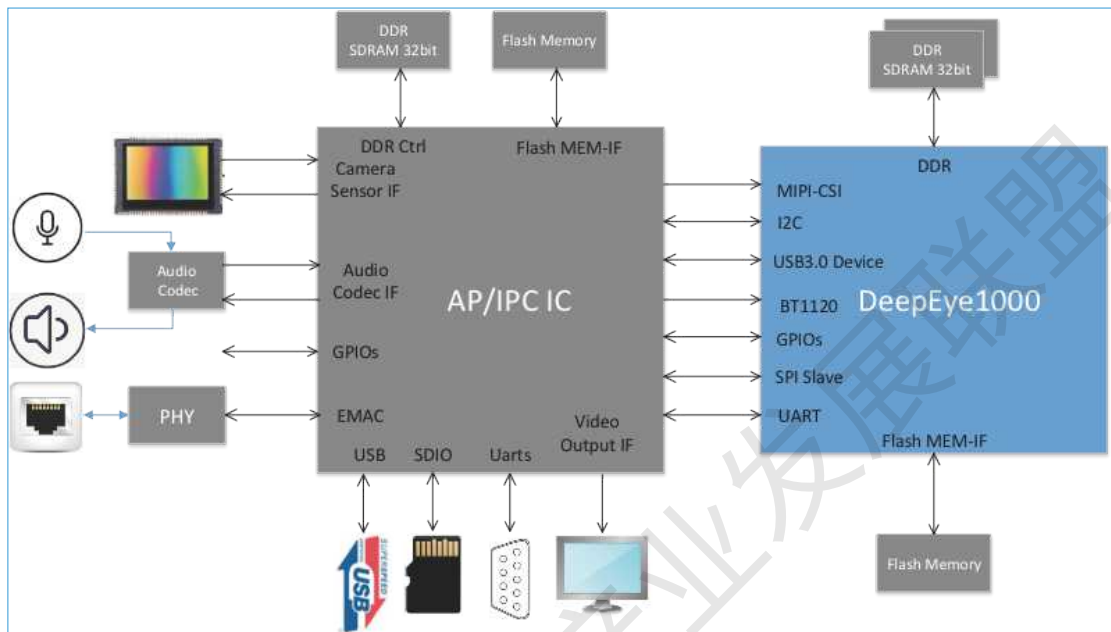


图 1 DeepEye1000 智能视觉协处理解决方案

说明：

- (1) DeepEye1000 作为协处理器协助主控芯片完成视频流或图像的 AI 处理能力；
- (2) 通过 MIPI, BT1120, USB3.0 接口接收来自主控端的实时视频流，完成对视频流的 AI 分析处理，把结果通过 USB 或 SPI 接口返回给主控芯片；
- (3) 支持最高 4K@30fps 实时视频流的视频分析处理；
- (4) 支持 YUV、RGB 和 RAW 格式的实时视频流输入；

2. DeepEye1000 智能视觉前置 AI 相机解决方案

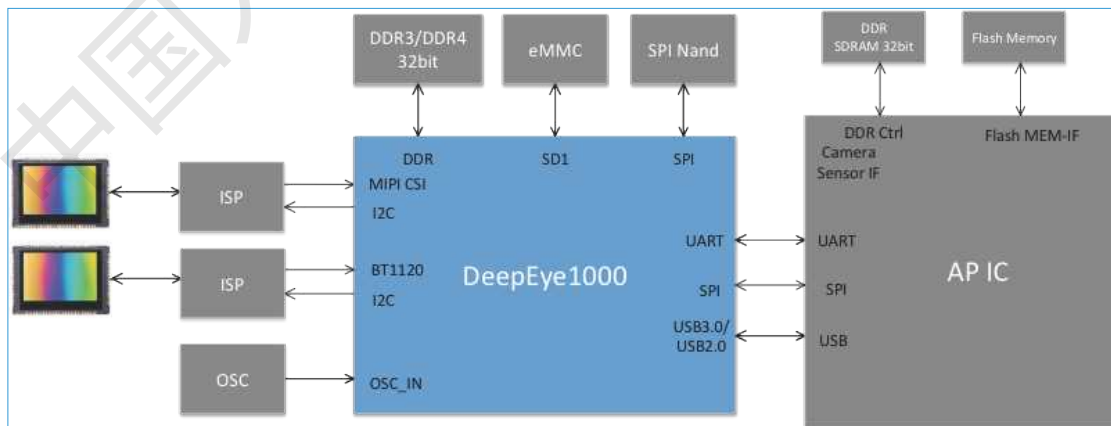


图 2 DeepEye1000 智能视觉前置 AI 相机解决方案

说明:

(1) DeepEye1000 通过 MIPI CSI 或 BT1120 接口灵活搭载前端专业成像系统,对输入成像视频流进行视觉 AI 全流程分析,完成视觉的前置化 AI 处理。处理结果通过 USB、UART 或 SPI 接口输出;

(2) DeepEye1000 通过 MIPI CSI 和 BT1120 搭载前端 3D 视觉前端,对输入深度点云信息和 RGB 信息,进行 3D 信息复原,3D 特征提取等,完成 3D 视觉的前置化 AI 处理。处理结果通过 USB、UART 或 SPI 接口输出;

(3) DeepEye1000 把算力前置,在图像传感器端完成实时的视觉 AI 分析,对主控 AP 选型几乎没有要求,可以搭载到任意主控 AP,极大的丰富了方案的灵活性;

(4) 芯片集成基于 2D/3D 生物识别算法,能被广泛用于人脸支付、商业零售、人脸门禁、智慧医疗、商业零售、AI 消费电子等领域;

(5) 芯片开发的工具链支持其他计算机视觉 CNN 算法移植和应用,搭配不同的前端成像系统,能被广泛用于如工业检测、机器人视觉、智慧医疗、智慧物流等领域。

3. 人脸识别模组方案

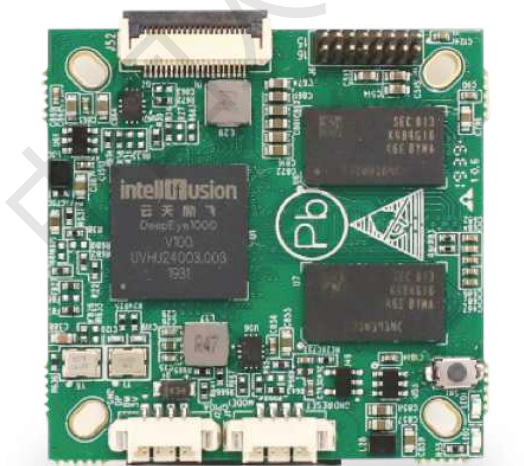


图 3 人脸识别模组方案

说明:

(1) 支持 200 万像素、1920×1080 视频分辨率

(2) 支持人脸检测、跟踪、选优、抓拍、识别、人脸属性分析

(3) 单场景可支持 40+ 人脸抓拍、20+ 人脸的本地特征提取和识别

(4) 最大支持 5 万人脸库离线识别,识别率大于 99%

(5) 支持远程人脸库管理功能

4. 多路视频并行处理 PCI-E 加速卡



图 4 多路视频并行处理 PCI-E 加速卡

说明:

(1) 单卡槽半高半长标准小卡,被动散热。

(2) 内置“云天初芯”芯片,支持 16 路 1080P 分辨率视频流的实时并行视频结构化分析,该方案提供 DeepEye1000 全栈工具链和 SDK,支持第三方算法模型部署,及上层应用开发。

(3) 适用于智能工业检测,边缘计算视频结构化加速等场景。

5. 边缘计算协处理方案



图 5 边缘计算协处理方案

说明：

(1) 支持多人像的动态抓拍、VIP 识别、人数统计、支持人脸、年龄、性别、人眼关注度等属性分析 可提供适配不同操作系统的主控 SDK，易于集成开发。

(2) 可广泛用于智慧门店场景下的精准营销、客流分析等。

【产品特征 / 技术特点】

1. 处理器内核

(1) 平头哥 32 位超高性能嵌入式 C810 CPU 玄铁处理器；

(2) 支持 16/32 位混合编码的 RISC 指令集；

(3) L1 I/D 32KByte, L2 128KByte, 主频最高 1.2GHz；

(4) 单位性能 2.5DMIPS/MHz；

(5) 支持大端和小端模式；

(6) 内部硬件调试模块支持片上硬件调试；

(7) 支持多电压、多时钟管理的低功耗技术；

2. 神经网络处理器

(1) 内含自研四核 NNP 神经网络处理器，主频最高 800Mhz；

(2) 自定义神经网络处理器指令集和架构；

(3) 支持 INT16/INT12/INT8 数据类型；

(4) 支持 CNN 算法所需的指令集；

(5) 支持 weight 参数压缩；

(6) 支持 feature map 压缩；

(7) 支持 caffe/tensorflow/mxnet 等主流深度学习框架；

度学习框架；

(8) 支持 Invasive 和 non-Invasive 的通用 Debug 架构；

(9) 支持处理器级联扩展、支持协处理器扩展；

(10) 时钟门控、电源门控、多电压、MVT 等典型低功耗技术；

3. DSP 处理器

(1) 双核 Vision Processor DSP，主频最高 600Mhz；

(2) 32KB I-Cache/64KB ITCM/256KB DTCM；

(3) 支持 OpenCV、OpenVX；

4. 视频编解码

(1) 支持多路 H.264、H.265 解码，性能为 4KP30；

(2) 支持 JTAG 编解码，其中解码支持最大 16K x 16K 分辨率，编码支持最大 8Kx8K 分辨率；

5. 视频输入接口

(1) 支持 MIPI CSI1.2 RX 4 Lane 输入接口，最高性能为 4KP30；

(2) 支持 BT1120 视频输入接口，性能为 1080P60；

(3) 支持 SDIO 3.0 或 USB 转 WIFI 接口接收视频；

(4) 支持 USB 转 EMAC 网口接收时视频；

6. 存储单元

(1) 支持 外接 DDR4/DDR3/DDR3L/LPDDR3 颗粒，最高速率 2677Mbps；

(2) 支持 SPI_NOR/SPI_NAND/EMMC/SD 等存储接口；

7. 外设接口

(1) 支持 USB3.0/2.0 Host/Device 接口；

(2) 支持 UART/GPIO/SPI/I2C 等通用的



低速接口；

8. 其它说明

- (1) 内含多个 Temp_sensor，支持功耗动态管理；
- (2) 支持 SPI_Nor Flash/SPI_Nand Flash SD/EMMC/USB/ 等多种 Boot 方式；
- (3) 支持版本在线升级。

【专利状态】

公司芯片专项发明专利近 200 件，并获得 2019 年度深圳市专利奖。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
DeepEye 1000 云天初芯	典型功耗 2W	支持最大 4GB	8GBps	2019 年 9 月	支持	支持	不支持	22nm

4.3.2 紫光展锐 – 虎贲 T710

【技术概述】

虎贲 T710 采用 8 核 CPU 架构，由 4 颗 2.0GHz 的 Arm Cortex-A75 及 4 颗 1.8GHz 的 Arm Cortex-A55 组成，搭载工作频率为 800MHz 的 IMG PowerVR GM 9446 图形处理器。虎贲 T710 包含了 CPU、GPU、NPU、ISP、VDSP 等处理单元，为各类丰富的 AI 应用提供了高效能、低功耗的技术基础。

【技术指标】

CPU: 4xA75@2.0GHz + 4xA55@1.8GHz
 GPU: PowerVR GM 9446@800MHz
 LCD: 支持 QHD+ 分辨率 (3200x1440)
 Camera: 支持 24M+8M/16M+16M 双摄
 NPU: 3.2TOPS@INT8 主频 800MHz
 网络制式: GSM/WCDMA/TDSCDMA/
 TDD-LTE/FDD-LTE
 WiFi: 802.11 a,b,g,n,11ac
 Bluetooth: 5.0

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps	2019 年 8 月	3.2 TOPS	0.5 TOPS		12nm

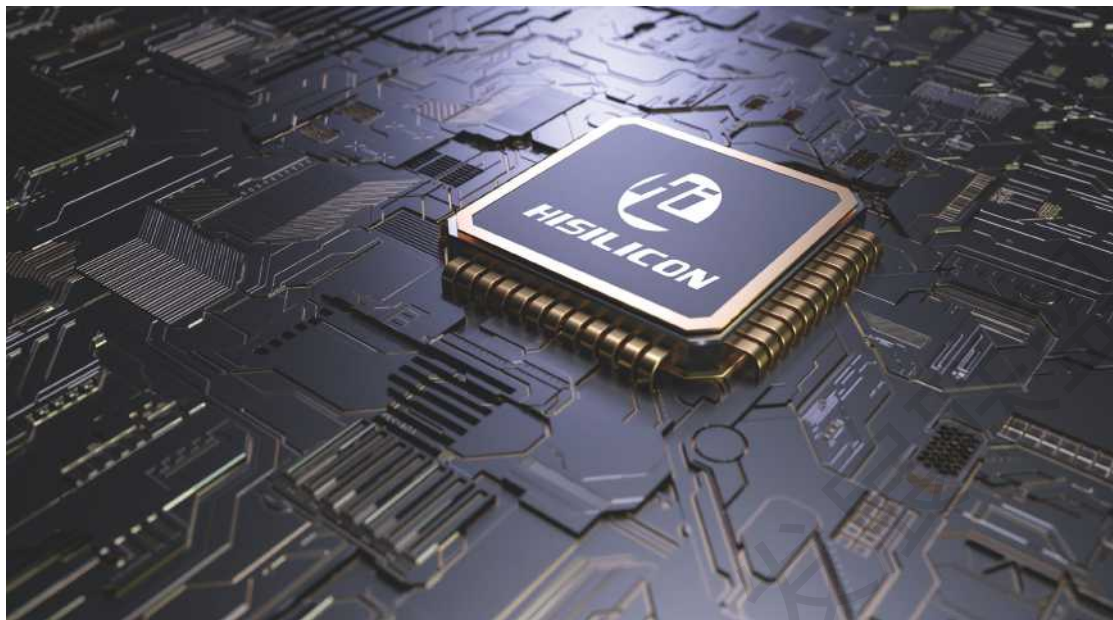


【应用案例】

虎贲 T710 已经在智能医疗，智能零售和智能安防多个领域内得到应用。

【产品特征 / 技术特点】

通过创新的异构双核架构，虎贲 T710 的计算能效比超过 2.5TOPS/W，并支持 INT4，INT8，INT16，FP16 等多种神经网络量化方式。同时还支持 4K@30 帧视频编解码，WiFi 11AC 和 BT 5.0 等多种无线连接技术。



4.3.3 上海海思技术有限公司 – Hi3516D V300

【技术概述】

Hi3516DV300 是新一代行业专用 Smart HD IP 摄像机 SOC，其宣称在低码率、高画质、智能处理和分析、低功耗等方面存在优势。

【技术指标】

Hi3516D V300 H.264/H.265 编解码最大宽度 2688，最大分辨率 2688x1944。

Hi3516D V300 集成神经网络加速引擎，处理性能达 1.0TOPS。

Hi3516D V300 提供 ISP 能力，提供多种图像处理能力。

【应用案例】

用于专业的端侧 IPC 相机、端侧人脸识别设

备（如门禁、考勤、支付终端等），具有强大的 ISP、编解码和智能平台能力。

【产品特征 / 技术特点】

Hi3516D V300 H.264/H.265 编解码最大宽度 2688，最大分辨率 2688x1944。

Hi3516D V300 集成了第四代 ISP，提供 WDR，多级 NR，6DoF DIS 以及多种图像增强和校正算法，从而确保了专业的图像质量。

Hi3516D V300 集成神经网络加速引擎，处理性能达 1.0TOPS，并集成智能计算加速引擎（含跟踪、人脸校正）。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3516D V300	典型 1.0W	外扩 2GB	32bit DDR3(L) DDR4	2018 年 10 月	1T			T28

4.3.4 上海海思技术有限公司 – Hi3519A V100



【技术概述】

Hi3519A V100 是一款高性能，低功耗 4K 智能 IP 摄像机 SoC。

【技术指标】

Hi3519A V100 引入了 H.265 / H.264 编码和解码，性能 4K x 2K @ 60 fps 和 1080p @ 240 fps。Hi3519A V100 集成了第四代 ISP，并提供低功耗工艺和架构设计。Hi3519A V100 提供共 2T 智能算力。

【应用案例】

用于专业的端侧 IPC 相机，具有强大的

ISP、编解码和智能平台能力。或用于边缘计算，用作多路 IPC 接入智能处理的中心。

【产品特征 / 技术特点】

Hi3519A V100 编解码性能 4K x 2K @ 60 fps 和 1080p @ 240 fps。

Hi3519A V100 集成了第四代 ISP，提供 WDR，多级 NR，6DoF DIS 以及多种图像增强和校正算法，从而确保了专业的图像质量。

Hi3519A V100 集成了 NNIE、DSP、CPU 等多核智能处理平台，提供 2T 智能算力。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3519A V100	典型 1.9W	外扩 4GB	DDR4/ DDR4 /LPDDR4/ SDRAM	2018 年 7 月	2T			T12



4.3.5 上海海思技术有限公司 – Hi3559A V100



【技术概述】

Hi3559A V100 是专业的 8K 超高清智能摄像机 SoC。

【技术指标】

它支持具有广播级图像质量的 8K30 / 4K120 数字视频录制，支持多个传感器输入以及 H.265 编码输出或电影级原始数据输出。它集成了高性能 ISP，并采用了低功耗工艺和架构设计。Hi3559A V100 提供出色的图像处理能力。Hi3559A V100 提供共 4T 智能算力。

【应用案例】

用于端侧 IPC 相机，具有 ISP、编解码和智能平台能力。或用于边缘计算，用作多路 IPC 接入智能处理的中心。

【产品信息】

芯片名称	功耗	内存	带宽		深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3559A V100	典型 3W	外扩 8GB	64-bitDDR4/ LPDDR4	2018 年 1 月	4T			T12

【产品特征 / 技术特点】

Hi3559A V100 支持多通道 4K 传感器输入，多通道 ISP 图像处理，HDR10 技术标准以及多通道全景硬件拼接。当支持 8K30 / 4K120 视频输入时，Hi3559A V100 提供了 6 轴 DIS。

Hi3559A V100 提供高效且丰富的计算资源，集成了双核 A73 处理器和双核 A53 处理器以及 big.LITTLE 架构和双操作系统，实现了功耗与启动时间之间的平衡。

Hi3559A V100 集成了 NNIE、DSP、CPU、GPU 等多核智能处理平台，提供 5T 智能算力。

【专利状态】（未提供）

4.3.6 上海海思技术有限公司 – Hi3562 V100

【技术概述】

Hi3562V100 是一颗面向车载行车记录仪、流媒体后视镜等领域推出的高性能、低功耗的 Camera SOC。

【技术指标】

Hi3562V100 支持 1080p30 H.265/H.264 编解码；集成了海思第四代 ISP，支持 WDR、多级降噪及多种图像增强和矫正算法，为客户提供专业级的图像质量；并提供 0.5TOPS 的神经网络加速器，以支持轻量智能业务。Hi3562V100 采用 28nm 低功耗工艺，符合 AECQ100 相关认证要求。

【应用案例】

用于车载行车记录仪、流媒体后视镜等领域。

【产品特征 / 技术特点】

低功耗，典型场景（1080p30 输入和 H.265 编码）功耗 0.6W。

提供最高 0.5T 的神经网络计算能力，提供神经网络加速，支持目标检测与识别等智能功能。

支持 2F-WDR，支持 1920x1080@30fps 2F-WDR 输入。

支持 MIPI 显示屏输出，支持流媒体后视镜等应用。

支持快速启动。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3562 V100	典型 0.6W	外扩 128MB~4GB DDR3L	16bit DDR3L	2020 年 3 月	0.5T			T28

4.3.7 上海海思技术有限公司 – Hi3566 V100

【技术概述】

Hi3566V100 是一颗面向车载行车记录仪、驾驶员状态监控等领域推出的高性能、低功耗的 Camera SOC。

【技术指标】

Hi3566V100 支持双路 1080p30 H.265/H.264 编码；集成了海思第四代 ISP，支持 WDR、多级降噪及多种图像增强和矫正算法，为客户提供专业级的图像质量；并提供 1TOPS 的神经网络加速器，以支持轻量智能业务。Hi3566V100 采用 28nm 低功耗工艺，符合 AECQ100 相关认证要求。

【专利状态】（未提供）

【产品信息】

【应用案例】

用于车载两路路输入，如行车记录仪、DMS 等场景。

【产品特征 / 技术特点】

支持双路 sensor 输入，支持双路 1920x1080@30fps 输入。

提供最高 1T 的神经网络计算能力，提供神经网络加速，支持目标检测与识别等智能功能。

支持 MIPI 显示屏输出，支持流媒体后视镜等应用。

低功耗，典型场景（双路 1080p30 输入和 H.264 编码）功耗 <1W。

支持快速启动。

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3566 V100	典型 <1W	外扩 256MB~4GB	32bit DDR3L	2020 年 3 月	1T			T28

4.3.8 上海海思技术有限公司 – Hi3568 V100

【技术概述】

Hi3568V100 是针对车载多路输入、高算力等需求设计的 SOC。

【技术指标】

Hi3568V100 它提供了 5 路输入，高性能 ISP 处理，支持 H.264/H.265 编码，为用户提供了卓越的图像处理能力。Hi3568V100 提供了高效且丰富的计算资源，支撑辅助驾驶类应用。Hi3568V100 集成了双核 A53，DSP 及高算力的深度学习硬件加速引擎。Hi3568V100 采用 FinFET 工艺，符合 AEC-Q100 认证要求。

【应用案例】

用于车载多路输入，ADAS 高算力需求的场

景，典型场景是多路环视、辅助泊车等。

【产品特征 / 技术特点】

支持多路视频输入，最大 5 路 1080p30 输入（2F line-mode WDR 融合后）。

神经网络运算能力，1.5TOPS INT8 运算能力。

互联扩展接口，支持 USB3.0 和 PCIe2.0 高速接口。

提供视觉计算处理能力，支持硬件深度信息处理，及其他视觉处理加速力。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3568 V100	典型 1.9W	外扩最大 4GB	LPDDR4	2020 年 6 月	1.5T			T12

4.3.9 上海海思技术有限公司 – Hi3569 V100

【技术概述】

Hi3569V100 是针对车载多路输入、高算力等需求设计的 SOC。

【技术指标】

Hi3569V100 提供了 8 路输入，高性能 ISP 处理，支持 H.264/H.265 编码，为用户提供了卓越的图像处理能力。

Hi3569V100 提供了高效且丰富的计算资源，支撑辅助驾驶类应用。Hi3569V100 集成了双核 A73 和双核 A53，DSP，GPU 及高算力的深度学习智能分析引擎。

Hi3569V100 采用 FinFET 工艺，符合 AEC-Q100 相关认证要求。

【应用案例】

用于车载多路输入、ADAS 高算力需求的场景。

【产品特征 / 技术特点】

支持多路视频输入。最大支持 8 路 1080 输入，且支持硬件拼接。

神经网络运算能力 4TOPS INT8 运算能力。

互联扩展接口。支持 USB3.0 和 PCIe2.0 高速接口。

提供视觉计算处理能力。支持硬件深度信息处理及其他视觉处理加速。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Hi3569 V100	典型 3W	外扩最大 8GB	64bit LPDDR4	2020 年 6 月	4T			T12

4.3.10 北京清微智能科技有限公司 – TX510

【技术概述】

可重构计算智能处理芯片，支持视觉、语音等多模态智能处理。

【技术指标】

- 1 典型功耗为 400mW，峰值算力达 1.2TOPS@INT8
- 2 支持混合精度计算和稀疏神经网络
- 3 AI 计算有效能效比达 5.6TOPS/W@INT8。

【应用案例】

人脸识别门锁：基于 3D 的活体分析、安防级人脸识别算法，达到亿分之一误识别（识别率 90%）的金融级安全，待机功耗小于 10uW，典型功耗小于 450mW，在平均每天开锁 20 次的极端情况下，8 节 5 号电池也可使用超过 1 年时间。启动时间小于 100ms+，识别时间小于 1 秒，免接触体验。

【产品特征 / 技术特点】

- 内置 3D 引擎（3D 结构光、TOF）

【产品信息】

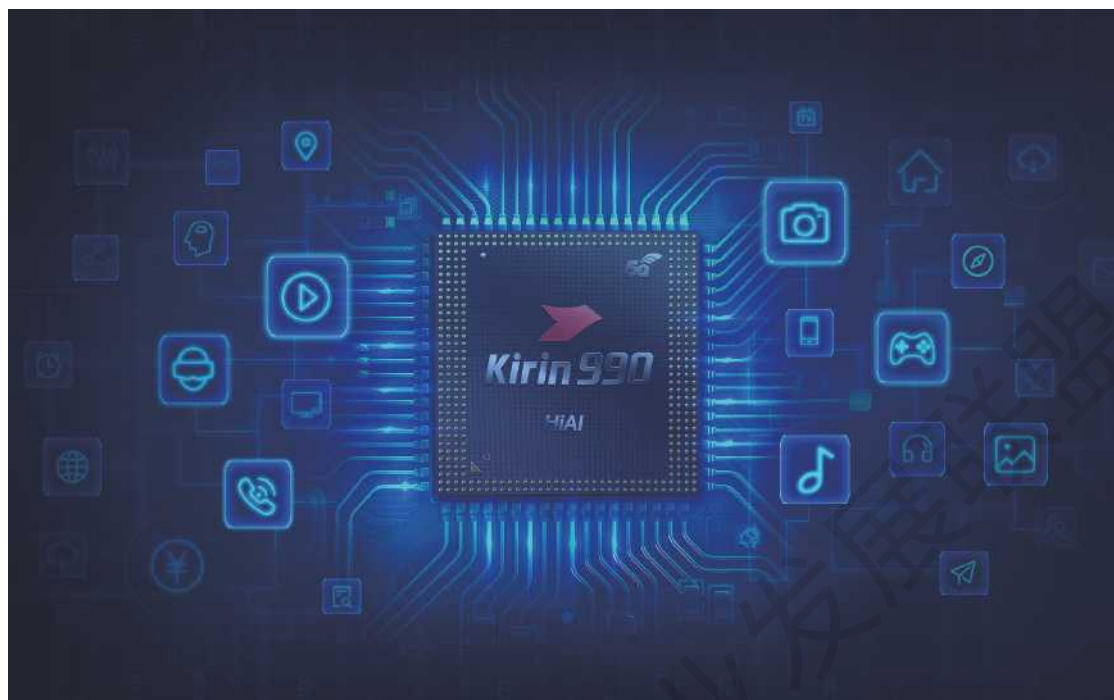


- 集成度高 (AIE、3DE、ISP、H264 等)
- 超低功耗 (待机功耗小于 10uW，峰值功耗小于 450mW)、AI 算力强 (1.2TOPS@INT8)，能效比高
- 芯片加密支持国密算法
- 开发简单，性价比高。

【专利状态】

申请发明专利 120 项，已获授权 55 项

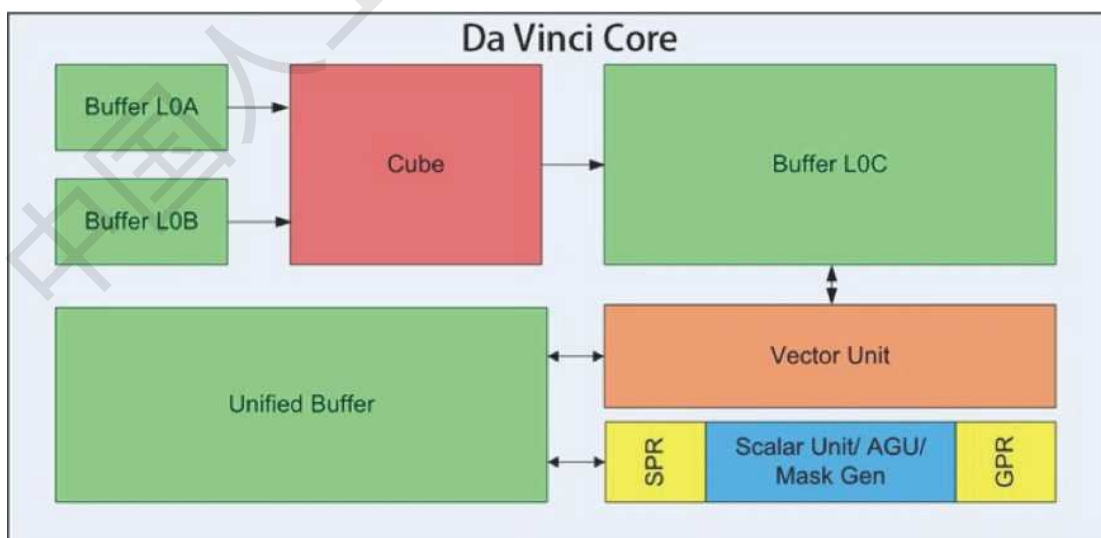
芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
TX510	工作模式 400mW	128MB	7.464GBytes/s	2020 年 7 月	1.2TOPS@600MHz			T C28HPC+



4.4 手机

4.4.1 华为技术有限公司 - Kirin 990 5G 处理器

【技术概述】



kirin 芯片中的 NPU 采用了华为的自研 NPU 架构，内部细分成很多单元，包括核心的 3D Cube、Vector 向量计算单元、Scalar 标量计算单元等，它们各自负责不同的运算任务实现并行化计算模型，共同保障 AI 计算的高效处理；矩阵乘是 AI 计算的核心，这部分运算由 3D Cube 完成。数据经过矩阵运算后，结果输出到 Buffer L0C 中，然后输送到 Vector 向量计算单元完成向量计算，类似 DSP，主要是池化（pooling）/padding、BN 运算、激活，加法等处理，如果得到最终的结果，会传输给 Unified Buffer。Scalar 标量运算单元负责标量运算，这里主要是前面 Cube、Vector 等指令的地址和参数计算以及基本的算术运算，同时还负责整个程序的循环控制和分支判断等工作，它的最终结果也会输出到 Unified Buffer。

目前 kirin NPU 可以支持超过 300 个 AI 算子数量，支持 Tensorflow/Caffe/Android NN 框架，也支持 IR 算子对接到第三方框架，在视觉计算神

经网络模型下，麒麟 990 5G 可以支持 90% 的开源模型。

【应用案例】

目前该芯片已经应用在华为多款手机中，并有许多智慧特性运行在 NPU 上，如：场景识别、美食分类，人脸检测，人脸识别，人像留色等应用；kirin NPU 通过 HiAI Foundation/HiAI Engine/HiAI Service 三层接口对外开放；目前已经开放 8 大类 30+ 个 Engine；

HiAI 已经与 15+ 应用厂商展开合作，让这些应用使能 kirin NPU 的能力，使得应用更加智能，增强应用的用户体验。

【产品特征 / 技术特点】

Kirin 990 5G 上，NPU 采用了 2 大核 +1 微核架构。

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
Kirin 990 5G 处理器	-	4x 16bit LPDDR4x @2133MHZ	DDR: 34.128 GB/S	2019 年 9 月	-	-	-	N7+

（Kirin 980 于 2019 年 3 月参与 AIIA DNN benchmark v0.5 评测。 ）

4.4.2 高通信息科技（上海）有限公司 – 高通骁龙™ 865 移动平台



【技术概述】

高通骁龙 865 移动平台是面向 5G 旗舰终端的最新一代平台。7nm 工艺，5G 调制解调器及射频系统可以提供 7.5 Gbps 的峰值速率，支持广网络覆盖、数据快速传输和全天电池续航。全新第五代 Qualcomm AI Engine 可实现每秒 15 万亿次运算。

【技术指标】

1.AI: 骁龙 865 全系统 AI，包括 CPU、

GPU、Hexagon 处理器、ISP、Qualcomm 传感器中枢（Sensing Hub）、安全处理单元、调制解调器，甚至 Quick Charge 等等。其 GPU 上改善了浮点运算和 AI 性能，将 TOPS 提升两倍，实现每秒 15 万亿次运算（15 TOPS）。同时，在 Hexagon 处理器上通过全新张量加速器将 TOPS 提升 4 倍。GPU 实现的 AI 计算最主要是 16 位和 32 位浮点运算，而 Hexagon 支持的是 8 位和 16 位的定点运算。同时开发的深度学习带宽压缩技术能够对这些海量数据进行 50% 的压缩，将需要进



入到芯片处理的数据压缩一半，以节省电能、降低功耗。这是一项完全无损的压缩技术。

2. 5G：骁龙 X55 5G 调制解调器及射频系统可支持 7.5 Gbp 的峰值速率。该 5G 全球解决方案支持所有关键地区和主要频段，包括毫米波以及 6GHz 以下 TDD 和 FDD 频段。其中，骁龙 865 能够在商用智能手机中实现毫米波支持，此外，它还支持 NSA 和 SA 组网模式、动态频谱共享（DSS）、全球 5G 漫游，并支持多 SIM 卡。支持实时双 Wi-Fi 6 技术。

3. 拍摄：骁龙 865 集成了 Spectra 480 ISP，支持以每秒 20 亿像素的速度处理专业品质的照片和视频。Spectra 480 ISP 处理速度支持骁龙 865 实现拍摄特性包括：8K 视频拍摄、杜比视界（Dolby Vision）视频拍摄、2 亿像素的照片拍摄、以及 960FPS 不限时慢动作视频拍摄。

4. 游戏：借助新一代 Snapdragon Elite

Gaming 特性，骁龙 865 提供移动游戏性能和端游级渲染品质，实现超流畅游戏体验。

【应用案例】

智能手机

【产品特征 / 技术特点】

- 1、从调制解调器到天线的完整 5G 解决方案
- 2、5G 毫米波
- 3、5G 6GHz 以下 200MHz 载波聚合
- 4、十亿像素级 ISP
- 5、拥有 15 TOPS 算力的第五代 AI Engine
- 6、持久稳定的图形性能
- 7、顶级游戏特性
- 8、实时双 Wi-Fi 6
- 9、蓝牙超宽带语音
- 10、支持 5G 全球部署的商用解决方案

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
高通骁龙™ 865 5G 移动平台	-	-	-	2019 年 12 月	✓	✓	✓	7nm

（注：骁龙 855 移动平台于 2019 年 6 月参与 AI1A DNN benchmark v0.5 测试  ）

4.5 语音

4.5.1 成都启英泰伦科技有限公司 – 人工智能语音识别降噪一体芯片 CI1103



【技术概述】

人工智能语音识别降噪一体芯片 CI1103 基于 NPU+MCU 架构设计，内置公司自主研发的神经网络处理器 BNPU，降噪等硬件算法模块，可实现噪声抑制、混响消除、回声消除、声纹识别、波束成形、声源定位、本地自学习等丰富功能。在应用上，该款芯片集成多路 UART、I2C、GPIO 等外围控制接口，可开发低成本的单芯片智能语音离线识别方案。同时，可外接 WIFI、蓝牙等无线模块，实现本地 + 云端语音方案。

【技术指标】

该款芯片主要有以下七项技术指标：

1. 自主研发神经网络处理器内核 BNPU，可独立进行语音信号的 DNN 神经网络计算，实现语音神经网络识别、端点检测等功能。该内核最大支持 6 个隐含层，每层 512 个神经元节点的运算，可以对 16KHz 的连续语音数据进行神经网络计算处理；

2. 自主研发降噪硬件算法模块，支持 1-2 路麦克风远场输入，可以实现噪声抑制、混响消除、回声消除等算法效果，最大支持 80dB 噪音抑制效果；

3. 支持 beamforming、DOA、本地自学习等算法；

4. 自主研发 UART、SPI、PWM 等芯片外设控制接口，可以实现最高 1Mbps 的串口通信，并可对接 Wi-Fi、BT、2.4G、433、zigbee 等通讯协议；

5. 自主研发多核 SOC 并行处理芯片架构，将神经网络处理器内核、降噪硬件算法模块和 ARM 的 Cortex-M4F 应用处理器内核及芯片外设控制接口结合，实现多核 SOC 单芯片，采用 55nm 工艺技术，系统主频为 160MHz；

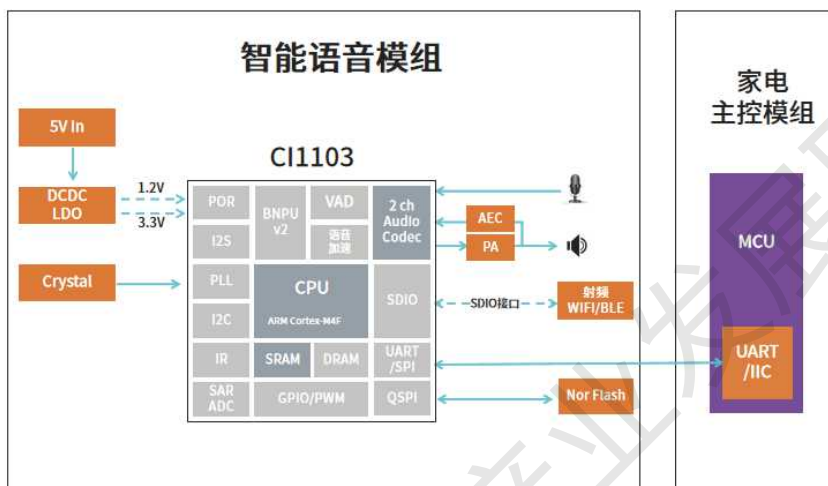
6. 开发芯片所需的智能语音算法，能实现本地 5 米远场语音识别功能，在环境噪音 40dB 时标准语音普通话输入时识别率可达 95%，声纹识别准确率大于 90%；

7. 该芯片可开发软硬件开发工具套件，并形成丰富的应用参考解决方案。

【应用案例】

1. “云 + 端” 智能家电解决方案

下图是应用框图：

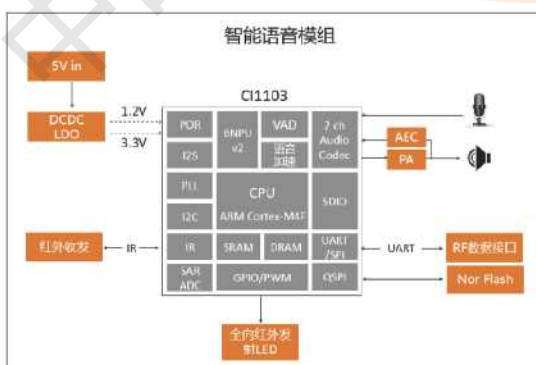


本应用方案支持 AEC（回声消除），可实现多种格式的流媒体音频播放并随时打断；可通过 SDIO 外接射频 WIFI/BLE 等，从而连接云端，将本地语音识别响应快、安全性高的优点和云端丰富资源的优点结合起来，并可进行 OTA 升级；外接 Nor Flash，可用于保存用户数据。

本方案可应用于油烟机、冰箱、微波炉等家电。

2. 智能语音红外解决方案

下图是应用框图：



工作流程是：人发出语音指令，通过麦克风将语音输入，智能语音模组进行识别，通过 UART/SPI 串口发送本地语音识别结果至家电主控模组，家电根据收到的信号开启不同的工作模式。

方案原理是：智能语音模组内置一个红外码库，每个红外信号有相对应的语音命令词，当语音模组接收到语音信号时，便转化为相应红外信号向外发射，实现对设备（带红外遥控功能）的控制。

该方案可应用于家居中控、开关面板、插座、家电单品等。比如“智能语音中控小艾”，通过红外学习，即可匹配灯具、电视机、空调、风扇、窗帘等家居设备，替换家中的手动红外遥控器，同时实现对多种设备的语音控制功能。

3. 声纹识别解决方案

声纹识别原理图如下：





操作流程是：用户首次使用时，说出唤醒词，系统检测并提取用户的语音特征，从而得到该用户的声纹模型。当用户再次使用时，说出唤醒词或命令词，系统即可匹配该声纹模型。该功能用于区分不同的角色，实现“听声识人”，让系统针对每个人提供不同的内容与服务。

该方案可应用于声纹锁、智能音箱、智能安防等产品。

【产品特征 / 技术特点】

人工智能语音识别降噪一体芯片 CI1103 是为 NPU+MCU 架构芯片，主要有以下六个特征：

1. 高算力：采用自主专利技术 2 代 BNPU 进行高性能神经网络并行计算，相当于数十个 CPU 并行计算能力；

2. 低功耗：基于 VAD 的低功耗芯片架构和特色低功耗设计，待机功耗最低只有 30mw，工作功耗 100mw；

3. 高集成低成本：集成 Audio Codec 模

块，硬件音频处理模块，多路 UART、I2C、SPI、PWM、GPIO、SDIO 等外围接口，极大降低了方案模组的硬件成本，模组成本约为同类 NPU+MCU 芯片的 1/2；

4. 快响应：本地离线识别，极速响应，最快仅需 0.2s，提升了用户体验感；

5. 多功能：支持本地语音识别（300 词条以上），本地 + 云端语音识别、中文（普通话、方言）/ 英文 / 日文识别、声纹识别、自然语义理解、语音合成、回声消除等；

6. 高可靠性：采用工业级设计，满足家电行业“双 85”标准要求（温度 85℃、湿度 85% 的环境下运行 22 天，实验结束后不能有任何品质问题）。

【专利状态】

公司基于人工智能语音识别降噪一体芯片 CI1103 的相关研究成果共申请了 86 篇专利，其中发明专利 38 篇，实用新型 40 篇，软件著作权 3 篇，集成电路布图 5 篇。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
人工智能语音识别降噪一体芯片 CI1103	待机功耗：30mw 工作功耗：100mw；	512KBSRAM； 2MBDRAM	1Gbps	2019 年 9 月	✓	✓	✓	55nm



4.5.2 上海深聪半导体有限责任公司 - 人工智能语音芯片

【技术概述】

深聪智能太行芯片是根据思必驰对语音交互算法的掌握及对市场的理解，自主定义开发的一款芯片 + 算法的人工智能人机语音交互的解决方案。可实现 AI 关键字和指令识别，低功耗唤醒。第一款芯片型号为 TH1520。

【技术指标】

在设计理念方面，TH1520 的主要关注点在于良好的用户体验和产品开发的灵活性。在设计方法方面，由于采取了软硬融合的方法，使得 TH1520 方案的性能和能效有了大幅度的提升。

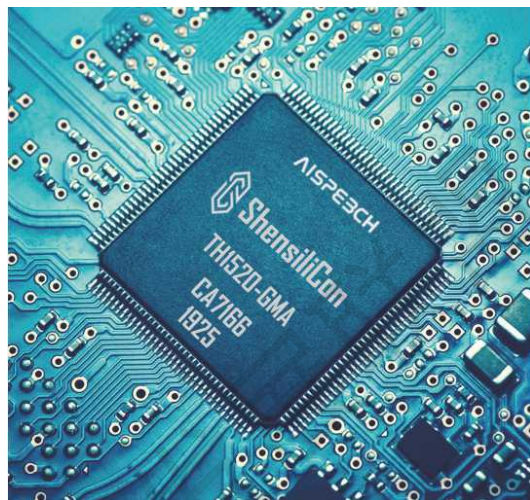
1. 低功耗：待机状态功耗毫瓦级，全速工作功耗在百毫瓦级。
2. 离线识别：支持 200 条以上指令的离线识别，可以全离线应用。
3. 片内存储：存储完全内置，节省成本和总体功耗。
4. 部署特点：提供算法 + 芯片的方案，支持快速将产品推向市场。
5. 适用性：多种接口，功耗模式，麦克风阵列灵活组合，适用各种场景。

【应用案例】

应用方向主要有智能家居、智能硬件终端、车载、手机、可穿戴设备等，聚焦于物联网语音芯片主战场。

目前第一款芯片已经用在智能家居场景，白电如智能空调，智能冰箱，智能洗衣机。黑电如智能电视，机顶盒。其他如晾衣架，取暖桌，Pad 等。

在疫情期间也提供了智能语音电梯解决方案，为非接触式场景提供了语音交互层面的解决方案。



【产品特征 / 技术特点】

1. 芯片特征

- (1) 双核增强型 DSP，含定制语音信号处理专用指令集
- (2) 专用的 AI 语音识别引擎
- (3) 大容量片内静态存储
- (4) 多通道音频编解码器，最高支持 6 + 2 通道语音同步采集
- (5) 支持全部标准音频格式，支持广域的采样率范围
- (6) 支持主流接口：USB/SPI/UART/I2C/I2S/GPIO
- (7) 可配置低功耗模式

2. 算法特征

- (1) 最高支持 6+2 通道语音同步采集，通道间延时小于 10ns
- (2) 语音端点检测，召回率：>99%，精度：>90%
- (3) 语音降噪，消除平稳和短时平稳噪声，SNR 增益：>15dB
- (4) 支持多达 6mic+2ref 的语音回声消除，

SNR 增益: >40dB

(5) 波束成形, 目标信号增强和干扰抑制,

SNR 增益: >15dB

(6) 语音侦听和目标语音唤醒, 召回率:

>95%, 误唤醒率: <1 次 /48 小时

(7) 声纹验证和声纹识别, 召回率 / 精度:

>95% / >97%

(8) 本地语音识别, 识别率 >95%

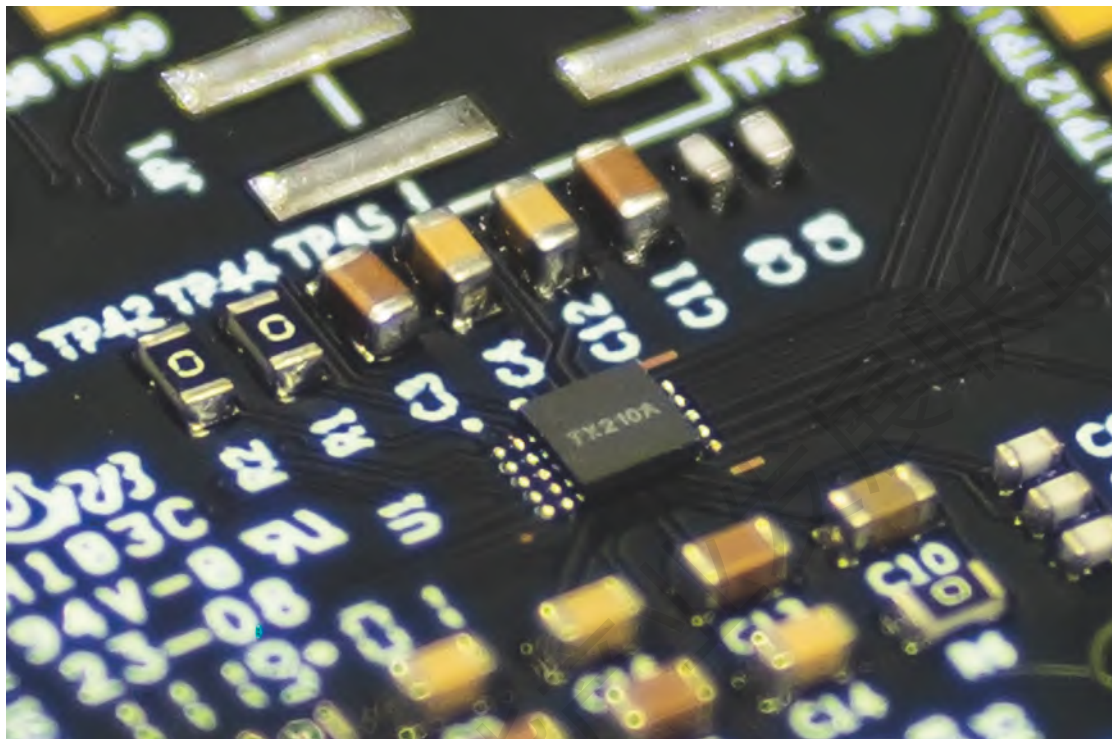
【专利状态】

集成电路布图证书 2 个, 软著申报 15 个, 其他专利申报中 6 个。

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
TH1520	200mW	3MB	3.2GB/S	2019 年 1 月	16GOPS	2GOPS	2GOPS	40nm

4.5.3 北京清微智能科技有限公司 – TX210



【技术概述】

采用可重构计算架构的智能语音芯片，适应于对功耗要求严苛的场景。

【技术指标】

- 1 工作功耗 450uW，VAD 功耗 200uW；
- 2 唤醒率超过 95%

【应用案例】

1 TWS 耳机：采用清微语音芯片 TX210，该耳机支持 always-on 语音唤醒，唤醒识别率 95%，误识别率小于 24 小时一次。TWS 耳机可保持长达 40 小时的待机时长。

2 智能家居：采用清微智能自主研发的

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
TX210	工作模式 450uW	700KB	3200 Mbps	2019 年 6 月	12.8GOPS @100MHZ			T 40nm

TX210 语音芯片，支持本地唤醒和离线命令词识别，识别率超过 95%。超长待机，可自由配置命令词列表，适用于不同的家居场景。

【产品特征 / 技术特点】

- 支持离线语音唤醒和多命令词识别
- 支持远场语音双麦降噪和回声消除
- 支持人声检测功能（VAD），降低长时间静默功耗
- 支持声纹识别，安全性高
- 支持一语直达，可以保存 3s 的语音数据

【专利状态】

申请发明专利 120 项，已获授权 55 项

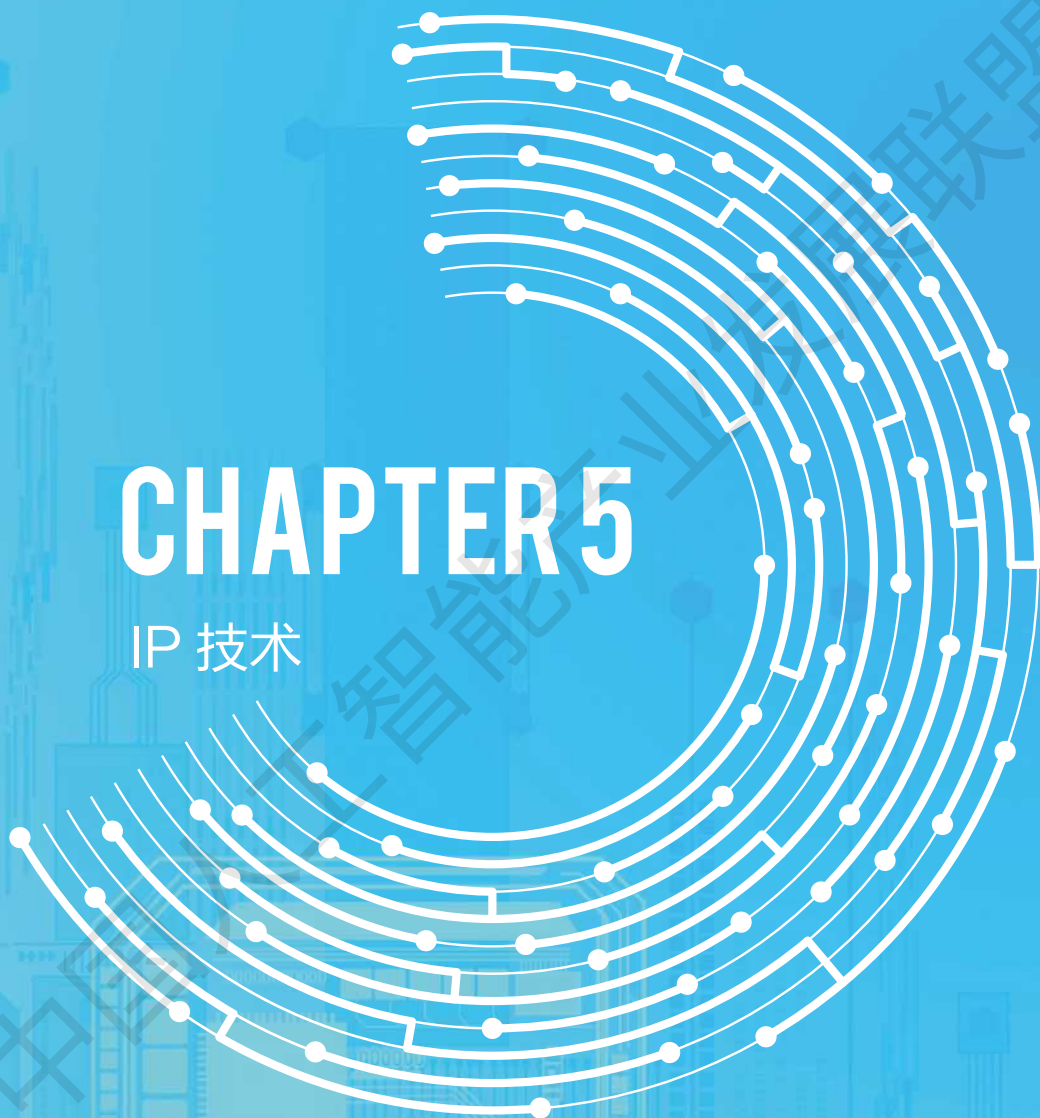
4.6 小结

终端	通用	公司名称	芯片名称	功耗	内存	带宽
		NVIDIA	Jetson Nano	5W 10W	4 GB64-bit LPDDR4LPDDR4	25.6GB/s
			Jetson TX2	7.5W 15W	8 GB128-bit LPDDR4	58.4GB/s
			Jetson Xavier NX	10W 15W	8 GB128-bit LPDDR4	51.2 GB/s
			Jetson AGX Xavier	10W 15W 30W	16GB 256-bit LPDDR4	137GB/s
			Jetson AGX Xavier	10W 20W	8GB 256-bit LPDDR4	85GB/s
		华为技术有限公司	昇腾 310AI 处理器	"8Tops@4W16Tops@8W"	LPDDR4X	2*64bit @3733MT/S
		瑞芯微电子股份有限公司	RK1808	3W	"LPDDR2/LPDDR3DDR3/DDR3L/DDR4"	6.4GBytes/s
			RK3399Pro	10W	双通道 DDR3-1866/DDR3L-1866/LPDDR3-1866/LPDDR4	14.9GBytes/s
	上海天数智芯科技有限公司	Iluvatar CoreX I	峰值 5W	8GB	273GB/s	
	智能驾驶	赛灵思	ZU7EV	10	/	/
			ZU11EG	16	/	/
		NVIDIA	DRIVE AGX Pegasus	300W	LPDDR4 + GDDR6	1 TB/s
			DRIVE AGX Xavier	30W	16 GB 256-bit LPDDR4	137 GB/s
	安防、机器人、车载	深圳云天励飞技术有限公司	DeepEye1000 云天初芯	典型功耗 2W	支持最大 4GB	8GBps
		紫光展锐	虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps
		上海海思技术有限公司	Hi3516D V300	典型 1.0W	外扩 2GB	"32bitDDR3(L)DDR4"
			Hi3519A V100	典型 1.9W	外扩 4GB	DDR4 / DDR4 / LPDDR4/SDRAM
			Hi3559A V100	典型 3W	外扩 8GB	"64-bit DDR4/LPDDR4"
			Hi3562 V100	典型 0.6W	外扩 128MB ~4GB DDR3L	16bit DDR3L
			Hi3566 V100	典型 <1W	外扩 256MB~4GB	32bit DDR3L
			Hi3568 V100	典型 1.9W	外扩最大 4GB	LPDDR4
			Hi3569 V100	典型 3W	外扩最大 8GB	64bit LPDDR4
		手机	北京清微智能科技有限公司	TX510	工作模式 400mW	128MB
	华为技术有限公司		Kirin 990 5G 处理器	-	4x 16bit LPDDR4x @2133MHZ	DDR: 34.128 GB/S
	高通信息科技（上海）有限公司		高通骁龙™865 移动平台			
语音	成都启英泰伦科技有限公司	人工智能语音识别降噪一体芯片 CI1103	待机功耗：30mW；工作功耗：100mW	512KBSRAM；2MBDRAM	1GB/s	
	上海深聪半导体有限责任公司	TH1520	200mW	3MB	3.2GB/s	
	北京清微智能科技有限公司	TX210	工作模式 450uW	700KB	3200Mbps	

上市时间	深度学习计算能力						工艺
	INT8	FP16	FP32	FP64	TF32	FP64 Tensor Core	
2019 年 3 月		0.5 TFLOPS					
2017 年 3 月		1.3TFLOPS					
2019 年 11 月	14 21TOPS						
2018 年 12 月	32TOPS	16TFLOPS					
2019 年	19.3TOPS	9.6TFLOPS					
2019 年	16TOPS	8T					12nm FFC
2019 年 6 月	3TOPS	100 GFLOPS	300 GOPS (INT16)				22nm
2018 年 1 月	3TOPS	0.3T	0.1T				28nm
2019 年 10 月	NA	4.8T	NA				T16
2019 年 11 月	/	/	/				16 nm UltraScale™
2019 年 11 月	/	/	/				16 nm UltraScale™
2018 年 11 月	320 TOPS						12nm
2018 年 11 月	30 TOPS						12nm
2019 年 9 月	支持	支持	不支持				22nm
2019 年 8 月	3.2TOPS	0.5TOPS					12nm
2018 年 10 月	1TOPS						T28
2018 年 7 月	2TOPS						T12
2018 年 1 月	4TOPS						T12
2020 年 3 月	0.5TOPS						T28
2020 年 3 月	1TOPS						T28
2020 年 6 月	1.5TOPS						T12
2020 年 6 月	4TOPS						T12
2020 年 7 月	1.2TOPS @600MHz						T C28HPC+
2019 年 9 月							N7+
2019 年 12 月	✓	✓	✓				7nm
2019 年 9 月	✓	✓	✓				55nm
2019 年 1 月	16GOPS	2GOPS	2GOPS				40nm
2019 年 6 月	12.8GOPS @100MHZ						T 40nm

CHAPTER 5

IP 技术



5.1 安谋科技（中国）有限公司 – 周易 AIPU



【技术概述】

Arm 中国自主研发的“周易”平台，是具有适配性强、开放通用等特点的人工智能平台，其主要包括全新的硬件处理器——人工智能处理单元（AIPU）。

【技术指标】

周易人工智能处理单元（AIPU）是一个专门用于深度学习处理的全新硬件 IP。这种新的架构采用了统一的一套指令集实现了 AI 计算的各种功能，包含张量指令（Tensor Instructions）、特定 AI 操作的指令以实现定制的硬件加速单元、以及 AI 计算的标量指令，同时支持用户自定义硬件实现。AIPU 可以实现人工智能应用的完全可编程性和很好的能效。除了 AIPU，还有完整的 AI 软件开发工具包，支持 AI 算法可以在 AIPU 上轻松部署。AIPU 采用了可配置的设计，包括张量执行单元（TEC）的数量和 AI 硬件加速单元，周易平

台能够涵盖从微控制器（100 GOPS）到高端边缘计算（十几 TOPS）的不同算力需求。AIPU 支持可选的安全扩展。

【应用案例】

基于周易 AIPU 设计的芯片面向 AI 的主流应用领域，如智能音箱 / 家电，智能门锁，智能蓝牙耳机 (TWS)，智能终端，安防，机器人，VR/AR，智能商业显示器，高级辅助驾驶 (ADAS) 等领域。

目前已有国内 AI 芯片头部企业基于周易 AIPU 设计的语音识别专业芯片（NPU）即将量产，为用户提供离线语音识别的强大能力，并大量应用于主流智能音箱品牌。

对于新开发的芯片，周易平台可以让芯片合作伙伴使用周易 AIPU IP 设计新的 SoC，并提供了验证测试向量、参考用例和后端实现参考流程等全面支持，以加速新的 SoC 设计；同时，周易



AIPU 仿真软件用于软件应用程序前期开发。

【产品特征 / 技术特点】

1. 创新架构

(1) 张量：新的架构设计了 AI 特定域张量指令集 (Tensor Instructions)

(2) 高能效比的专用硬件加速：特定 AI 操作的指令集以实现定制的硬件加速单元标量

(3) 标量：用于 NN 计算所必需的通用标量指令集

2. PPA 最佳平衡

(1) 可扩展性配置

(2) 高密度性能

3. 可选的安全扩展：有效保护用户信息，AI 算法

4. 完整工具链

(1) Build 工具和 C 编译器（离线）：模型转换和优化，生成周易可执行文件（一键生成）

(2) 运行库：Driver and Firmware

(3) 仿真器 和 调试器

5. 支持主流 AI 框架 (TensorFlow/ TensorFlow Lite/ONNX 等)

6. 丰富的 AI 应用：基于 CPU/GPU/AIPU 的应用优化；开发者社区

【专利状态】（未提供）

【产品信息】

芯片名称	功耗	内存	带宽	上市时间	深度学习计算能力			工艺
					INT8	FP16	FP32	
周易 AIPU	>2TOPS/W@28nm	Configurable Default 256KB	2~2.5 GB / TOPS	2019 年 6 月	0.2 - 4TOPS	n/a	n/a	28nm

【补充信息】

周易 AIPU 是一个硬件的 AI IP, 客户可以给予周易 AI IP 设计自己的覆盖 AI 全场景的 AI 芯片。

5.2 颖脉信息技术（上海）有限公司 – PowerVR NNA 神经网络加速器系列 IP



(mobile phone SOC UNISOC T710 )

【技术概述】

PowerVR NNA 是 Imagination 公司推出的完整、独立式的系统性硬件 IP 神经网络处理器，是一款具备支持从软件、硬件、工具全套的神经网络系统解决方案，同时提供 16 位到 4 位高灵活度位宽的灵活调整能力。其架构可支持多种操作系统，包括 Linux 和 Android 等等。目前最新版本为第三代 PowerVR 3NX NNA 神经网络加速器 IP。

【技术指标】

Imagination 为开发人员提供所有必要的工具，可快速、轻松地启用和执行其神经网络，并确保运算带宽能与准确度平衡。PowerVR 2NX 具有移动 Android 领域的完整支持，其开发资源包括映射 (mapping) 和微调工具、样本网络、评估工

具与文件。PowerVR NX Mapping Tool 可实现机器学习框架包括 Caffe 和 Tensorflow 的转换。

Series3NX 提供了可扩展性，使 SoC 制造商能够优化汽车，移动，智能监控和物联网边缘设备等嵌入式市场的计算能力和性能。同时也带来了一些架构上改进，包括无损压缩技术，Series3NX 架构在同一硅片面积上性能比前一代产品提高了 40%，使 SoC 制造商的性能效率提高了近 60%，带宽降低了 35%。

基于 PowerVR 2NX IP 上市的紫光展锐虎贲 T710 芯片

计算能效比超过 2.5TOPS/W，并支持 INT4，INT8，INT16，FP16 等多种神经网络量化方式。同时还支持 4K@30 帧视频编解码，WiFi 11AC 和

BT 5.0 等多种无线连接技术。

基于 PowerVR 3NX IP 技术的芯片将于 2020 年下半年 tape out。

【应用案例】

海信 5G 手机 F50 采用紫光展锐虎贲 T7510 芯片，该机采用了紫光展锐虎贲 T7510，其内部搭载了虎贲 T710 处理器 + 春藤 V510 调制解调器。T710 图形方面采用了 Imagination PowerVR GM 9446 GPU，主频 800Hz，满足各种使用场景的性能需求。在 AI 方面也采用了 Imagination PowerVR 2NX NNA 神经网络加速器 IP。内置的各种处理单元，能够提供出色的 AI 性能，为各种 AI 场景提供了高效的性能支持。

【产品特征 / 技术特点】

可提供：

1. 灵活的位宽以及数据类型支持

作为一种完全灵活的解决方案，Series3NX 支持从 16 位到 4 位的数据位宽，换取性能的提升和带宽资源的节省。

2. 参数的无损压缩

作为对可配置数据位宽特性的补充，Series3NX 引入了无损权重压缩方案，可减少网络模型大小并节省带宽，从而提高整体性能。

3. 高级安全性支持

PowerVR Series3NX 集成了通用的安全方案，灵活的基础架构使其可以被方便地集成到用户的系统级解决方案中，使数据权限持有者能够在需要时保护其内容。

【专利状态】

Imagination 专利技术

【产品信息】

架构：PowerVR Series2NX

Series2NX 提供了市场上高推理 / mm²，可创建紧凑且经济高效的推理解决方案。

(1) 可扩展架构

(2) 灵活的位宽支持：作为一种完全灵活的解决方案，Series2NX 支持从 16 位到 4 位的神经网络位宽，从而在不影响推理精度的情况下减少了带宽并提高了性能。

(3) 低功耗：凭借业界领先的推理 / mW，

Series2NX 以低功耗提供一流的神经网络加速。

架构：PowerVR Series3NX

PowerVR Series3NX 是市场上用于神经网络硬件加速的嵌入式解决方案。由于关键的架构增强，3NX 系列的性能比上一代产品提高了 40%，单核每秒可执行 10tera 的操作 (TOPS)，提供了市场上每平方毫米超高的性能密度。

Model	Engines	8-bit TOPs	16-bit TOPs	8-bit MACs	16-bit MACs	APIs
AX2145	1	1 TOPs	0.5 TOPs	512 MACs/clock	256 MACs/clock	IMGDNNAPI Android NN HAL
AX2185	8	4.1 TOPs	2.0 TOPs	2048 MACs/clock	1024 MACs/clock	IMG DNN API Android NN HAL

(1) 灵活的位宽支持

3NX 系列产品支持从 16 到 4 位宽的神经网络，可减少带宽和提高性能而不影响推理精度。

(2) 无损压缩

作为对低位宽支持的补充，Series3NX 引入了一种无损权重压缩方案，它减少了网络模型的大小和带宽，从而提高了整体性能。

(3) 安全支持

PowerVR 3NX 系列集成了安全架构，包括灵活的底层基础，使其集成到自定义解决方案，确

保权利持有人保护他们的内容在需要的地方。

(4) 低功耗

凭借业界高推理 /mW，Series3NX 以低功耗提供一流的神经网络加速。

PowerVR Series3NX Multi-Core Option

(3NX 系列多核选择)

由于 PowerVR Series3NX 架构的可扩展性，多核可最多实现 160 个 TOPs，。Series3NX 将提供多种产品，使 SoC 制造商能够满足一系列设计目标，以解决多个市场和应用程序。

Model	8-bit TOPs	8-bit MACs	16-bit MACs	APIs
AX3125	0.6 TOPS	256 MACs/clock	64 MACs/clock	IMG DNN APIAndroid NN HAL
AX3145	1.2 TOPS	512MACs/clock	128MACs/clock	IMG DNN APIAndroid NN HAL
AX3365	2.0	1024	256	IMG DNN APIAndroid NN HAL
AX3385	4.0	2048	512	IMG DNN APIAndroid NN HAL
AX3595	10.0	4096	1024	IMG DNN APIAndroid NN HAL

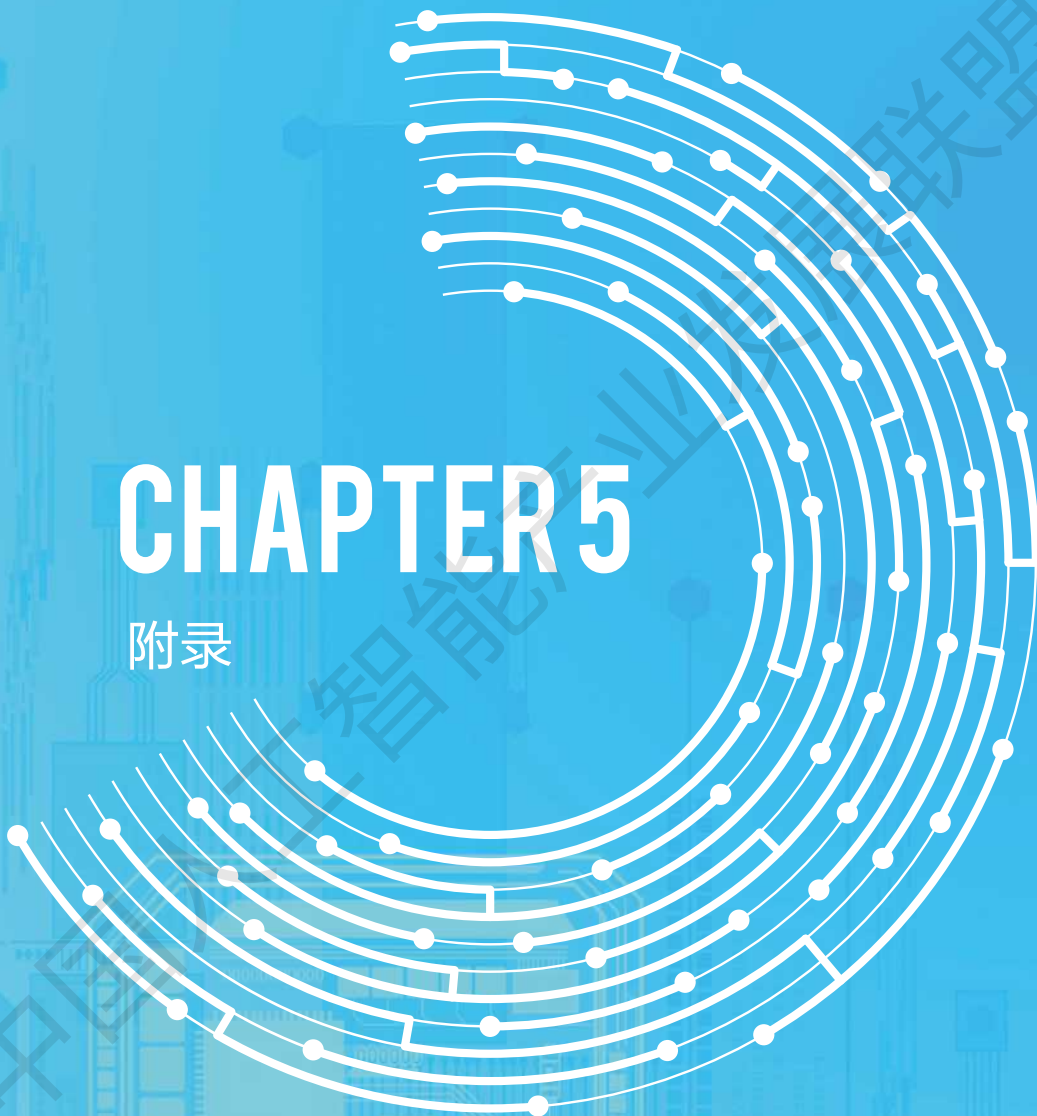
Model	Cores	8-bit TOPs	8-bit MACs	16-bit MACs	APIs
UH2x40	2	20	8192	2048	IMG DNNAPI Android NN HAL
UH4x40	4	40	16384	4096	IMG DNNAPI Android NN HAL
UH8x40	8	80	32768	8192	IMG DNNAPI Android NN HAL
UH16x40	16	160	65536	16384	IMG DNNAPI Android NN HAL

5.3 小结

IP 技术	公司 名称	芯片 名称	功耗	内存	带宽	上市 时间	深度学习计算能力						工艺
							INT8	FP16	FP32	FP64	TF32	FP64 Tensor Core	
	安谋科技 (中国) 有限公司	周易 AIPU	>2TOPS /W@28nm	Configurable Default 256KB	2~2.5 GB / TOPS	2019 年 6 月	0.2 - 4TOPS	n/a	n/a				28nm
	颖脉信息技术 (上海) 有限公司	包含 3 种设计架构: PowerVR Series2NX, PowerVR Series3NX 和 PowerVR Series3NX Multi-Core Option 架构											

CHAPTER 5

附录



AI 芯片选型目录 (2020 年 7 月版)

云端 / 数据中心	云端训练	公司名称	芯片名称	功耗	内存	带宽	上市时间
		北京百度网讯科技有限公司	百度昆仑 1	150W	16GB	512GB/s	2019 年 11 月
		上海燧原科技有限公司	邃思	190W(最大功耗)	16GB HBM	512GB /s	2020 年 6 月
		NVIDIA	NVIDIA A100	400W	40GB HBM2	1555 GB/s	2020 年 5 月
	NVIDIA V100(NVLink)		300W	32/16 GB HBM2	900 GB/s	2017 年 5 月	
	NVIDIA V100 (PCIe)		250	32/16 GB HBM2	900 GB/s	2017 年 5 月	
	NVIDIA V100S		250W	32 GB HBM2	1134 GB/s	2019 年 11 月	
	NVIDIA	NVIDIA T4	70W	16 GB GDDR6	320+ GB/s	2018 年 3 月	
	云端推断	赛灵思	Alveo U50 (VU35P)	50W	8GB	316GB/s	2019 年 8 月
			Alveo U200 (VU9P)	100W	64GB	77GB/s	2018 年 10 月
Alveo U250 (VU13P)			110W	64GB	77GB/s	2018 年 10 月	
Alveo U280 (VU37P)			100W	/	/	2018 年 11 月	
中科寒武纪科技股份有限公司			思元 100	75W	8GB/16GB	102GB/s	2018 年 5 月
北京比特大陆科技有限公司		BM1684	16W	12GB LPDDR4X	68.3GB/s	2019 年 9 月	
深圳鲲云信息科技有限公司	星空加速卡 X3	峰值 56W	两组 DDR4SDRAM, 8GB/DDR	32GB/s	2020 年 6 月		
华为技术有限公司	昇腾 310AI 处理器	"8Tops @4W16Tops@8W"	LPDDR4X	2*64bit @3733MT/S	2019 年		
边缘计算	上海天数智芯科技有限公司	Iluvatar CoreX I	峰值 5W	8GB	273Gb/s	2019 年 10 月	
	NVIDIA	Jetson Nano	5W 10W	4 GB64 -bit LPDDR4	25.6GB/s	2019 年 3 月	
		Jetson TX2	7.5W 15W	8 GB128 -bit LPDDR4	58.4GB/s	2017 年 3 月	
		Jetson Xavier NX	10W 15W	8 GB128 -bit LPDDR4	51.2 GB/s	2019 年 11 月	
		Jetson AGX Xavier	10W 15W 30W	16GB 256 -bit LPDDR4	137GB/s	2018 年 12 月	
		Jetson AGX Xavier	10W 20W	8GB 256 -bit LPDDR4	85GB/s	2019 年	
		NVIDIA T4	70W	16 GB GDDR6	320+ GB/s	2018 年 3 月	
	紫光展锐	虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps	2019 年 8 月	
	中科寒武纪科技股份有限公司	思元 220 M.2	8.25W	4GB	29GB/s	2019 年 11 月	
	华为技术有限公司	昇腾 310AI 处理器	"8Tops @4W16Tops @8W"	LPDDR4X	2*64bit @3733MT/S	2019 年	
	深圳鲲云信息科技有限公司	星空加速卡 X3	峰值 56W	两组 DDR4 SDRAM, 8GB/DDR	32GB/s	2020 年 6 月	

深度学习计算能力						工艺
INT8	FP16	FP32	FP64	TF32	FP64 Tensor Core	
256TOPS	64 TOPS	16TOPS				14nm
80TOPS	80TFLOPS	20TFLOPS				GF 12nm FinFET
"624 TOPS 1248 TOPS* (* 采用稀疏技术) "	312 TF 624 TF* (* 采用稀疏技术)	19.5 TFOPS	9.7 TFOPS	"312 TFOPS (* 采用稀疏 技术) "	19.5 TFOPS	7nm
	125TFLOPS	15.7TFLOPS	7.8 TFLOPS			12nm
	112TFLOPS	14 TFLOPS	7TFLOPS			12nm
	130TFLOPS	16.4TFLOPS	8.2TFLOPS			12nm
130TOPS	65TFLOPS	8.1TFLOPS				12nm
16.2TOPS	/	/				16nm UltraScale™
18.6TOPS	/	/				16nm UltraScale™
33.3TOPS	/	/				16nm UltraScale™
24.5TOPS	/	/				16nm UltraScale™
32TOPS	16TFLOPS	/				16nm
128 TOPS	/	/				16nm
"17.6TOPS (普通模式) 35.2TOPS (Winograd 加速模式) "		2.2T				12nm
10.9 TOPS	N/A	N/A				28nm
16TOPS	8T					12nm FFC
NA	4.8TFLOPS	NA				T16
	0.5 TFLOPS					
	1.3TFLOPS					
14 21TOPS						
32TOPS	16TFLOPS					
19.3TOPS	9.6TFLOPS					
130TOPS		8.1TFLOPS				12nm
3.2TOPS	0.5TOPS					12nm
8TOPS	/	/				16nm
16TOPS	8T					12nm FFC
10.9 TOPS	N/A	N/A				28nm

终端	通用	NVIDIA	Jetson Nano	5W 10W	4 GB64-bit LPDDR4	25.6GB/s	2019 年 3 月	
			Jetson TX2	7.5W 15W	8 GB128-bit LPDDR4	58.4GB/s	2017 年 3 月	
			Jetson Xavier NX	10W 15W	8 GB128-bit LPDDR4	51.2 GB/s	2019 年 11 月	
			Jetson AGX Xavier	10W 15W 30W	16GB 256-bit LPDDR4	137GB/s	2018 年 12 月	
			Jetson AGX Xavier	10W 20W	8GB 256-bit LPDDR4	85GB/s	2019 年	
		华为技术有限公司	昇腾 310AI 处理器	"8Tops @4W16Tops @8W"	LPDDR4X	2*64bit @3733MT/S	2019 年	
		瑞芯微电子股份有限公司	RK1808	3W	"LPDDR2 /LPDDR3DDR3/ DDR3L/DDR4"	6.4GB ytes/s	2019 年 6 月	
			RK3399Pro	10W	双通道 DDR3-1866/DDR3L -1866/LPDDR3 -1866/LPDDR4	14.9GB ytes/s	2018 年 1 月	
	上海天数智芯科技有限公司	lluvatar CoreX I	峰值 5W	8GB	273GB/s	2019 年 10 月		
	智能驾驶	赛灵思	ZU7EV	10	/	/	2019 年 11 月	
			ZU11EG	16	/	/	2019 年 11 月	
		NVIDIA	DRIVE AGX Pegasus	300W	LPDDR4 + GDDR6	1 TB/s	2018 年 11 月	
			DRIVE AGX Xavier	30W	16 GB 256-bit LPDDR4	137 GB/s	2018 年 11 月	
	安防、机器人、车载	深圳云天励飞技术有限公司	DeepEye1000 云 天 初 芯	典型功耗 2W	支持最大 4GB	8GBps	2019 年 9 月	
		紫光展锐	虎贲 T710	<2.3W	LPDDR4X 1866	15Gbps	2019 年 8 月	
		上海海思技术有限公司	Hi3516D V300	典型 1.0W	外扩 2GB	"32bitDDR3 (L)DDR4"	2018 年 10 月	
			Hi3519A V100	典型 1.9W	外扩 4GB	DDR4/ DDR4/ LPDDR4/SDRAM	2018 年 7 月	
			Hi3559A V100	典型 3W	外扩 8GB	"64-bit DDR4 /LPDD R4"	2018 年 1 月	
			Hi3562 V100	典型 0.6W	外扩 128MB ~4GB DDR3L	16bit DDR3L	2020 年 3 月	
			Hi3566 V100	典型 <1W	外扩 256MB~4GB	32bit DDR3L	2020 年 3 月	
			Hi3568 V100	典型 1.9W	外扩最大 4GB	LPDDR4	2020 年 6 月	
			Hi3569 V100	典型 3W	外扩最大 8GB	64bit LPDDR4	2020 年 6 月	
		北京清微智能科技有限公司	TX510	工作模式 400mW	128MB	7.464GBytes/s	2020 年 7 月	
		手机	华为技术有限公司	Kirin 990 5G 处理器	-	4x 16bit LPDDR4x @2133MHZ	DDR: 34.128 GB/S	2019 年 9 月
			高通信息科技（上海）有限公司	高通骁龙™ 865 移动平台				2019 年 12 月
		语音	成都启英泰伦科技有限公司	人工智能语音识别 降噪一体芯片 C11103	待机功耗：30mW； 工作功耗：100mW	512KBSRAM； 2MBDRAM	1GB/s	2019 年 9 月
			上海深聪半导体有限责任公司	TH1520	200mW	3MB	3.2GB/s	2019 年 1 月
			北京清微智能科技有限公司	TX210	工作模式 450uW	700KB	3200Mbps	2019 年 6 月
IP 技术		安谋科技（中国）有限公司	周易 AIPU	>2TOPS/ W@28nm	Configurable Default 256KB	2~2.5 GB / TOPS	2019 年 6 月	
	颖脉信息技术（上海）有限公司	包含 3 种设计架构：PowerVR Series2NX， PowerVR Series3NX 和 PowerVR Series3NX Multi-Core Option 架构						

考研资料：数学、英语、政治、管综、西综、法硕等（整合各大机构）

英语类：四六级万词班专四专八雅思等

财经类：初级会计、中级会计、注册会计师、税务师、会计实操、证券从业、基金从业、资产评估、初级审

公务员：国考、省考、事业单位、军队文职、三支一扶微信 2270291360

银行：银行招聘、笔试、面试

教师资格：小学、中学、教师招聘面试

建筑：一建、二建、消防、造价

法考：主观题、客观题

多平台网课：涵盖职场、办公技能、编程、文案写作、情感心理、穿搭技巧、理财投资健身减肥摄影技术等优质内容

精选资料：Excel 教程、PPT 模板、简历模板、PS 教程、PPT 教程、素描、烹饪、小语种、CAD 教程、PR 教程、UI

课程、自媒体、写作、计算机二级、钢琴、Python、书法、吉他、kindle 电子书、演讲.....持续更新中...

押题：提供考前冲刺押题（初级会计、中级会计、注册会计师、一建、二建、教资、四六级、证券、基金、期货等等），麻麻再也不用担心我考不过了。

资料领取微信：1131084518

行业报告：20000 份+持续更新

英语四六级备考资料	计算机二级备考资料	150 所高校考研专业课资料
两小时搞定毛概马原思修近代史纲	高数(微积分)+线性代数+概率论	素描 0 基础入门教程
教师资格证全套备考资料	普通话考试资料礼包	书法教程微信 2270291360
大学生英语竞赛备考资料	大学生数学竞赛备考资料	1000 份各行业营销策划方案合集
挑战杯/创青春/互联网+竞赛资料	电子设计竞赛必备资料	街舞 0 基础入门教程
托福雅思备考资料	大学物理学科攻略合集	动漫自学教程
SCI 最全写作攻略	TEM4/TEM8 专四专八备考资料	教师资格证面试试讲万能模板
360 份精美简历模板	数学建模 0 基础从入门到精通	100 套快闪 PPT 模板
Vlog 制作最全攻略	超强 PR 模板	42 套卡通风 PPT 模板
PS 零基础教程微信 1131084518	PS 高级技能教程	63 套酷炫科技 PPT 模板
好用到极致的 PPT 素材	128 套中国风 PPT 模板	32 套 MBE 风格 PPT 模板
327 套水彩风 PPT 模板	295 套手绘风 PPT 模板	54 套毕业答辩专属 PPT
196 套日系和风 PPT 模板	82 套文艺清新 PPT 模板	57 套思维导图 PPT 模板
163 套学术答辩 PPT 模板	53 套北欧风 PPT 模板	34 套温暖治愈系 PPT 模板
118 套国潮风 PPT 模板	30 套仙系古风 PPT	126 套黑板风 PPT 模板
114 套星空风格 PPT 模板	192 套欧美商务风 PPT 模板	42 套绚丽晕染风 PPT
50 套精美 INS 风 PPT 模板	56 套水墨风 PPT 模板	137 套清爽夏日风 PPT 模板
98 套森系 PPT 模板	25 套简约通用 PPT 模板	记忆力训练教程
300 套教学说课 PPT 模板	123 套医学护理 PPT 模板	AE 动态模板微信 2270291360
毕业论文资料礼包	教师资格证重点笔记+易错题集	表情包制作教程
吉他自学教程（送 6000 谱）	钢琴自学教程（送 1000 谱）	区块链从入门到精通资料
2000 部 TED 演讲视频合集	Excel 从入门到精通自学教程	单片机教程
230 套可视化 Excel 模板	1000 款 PR 预设+音效	1000 份实习报告模板
手绘自学教程微信 1131084518	单反从入门到精通教程	人力资源管理师备考资料
英语口语自学攻略	粤语 0 基础从入门到精通教程	证券从业资格考试备考资料
日语自学教程	韩语自学教程	PHP 从入门到精通教程
法语学习资料	西班牙语学习资料	炒股+投资理财从入门到精通教程
全国翻译专业资格考试备考资料	BEC 初级+中级+高级全套备考资料	大数据学习资料
SPSS 自学必备教程	Origin 自学必备教程	会计实操资料
LaTeX 全套教程+模板	EndNote 教程+模板	小提琴 0 基础入门自学教程
GRE 超全备考资料	200 份医学习题合集	司考备考资料

上万 GB 教学资料 (均全套, 非杂乱) 免费领取微信 1131084518

《闪电式百万富翁》实战版+升级版

易经+道德经+易学名师全集+风水学+算命学+起名+++等等 (全套 1000 多 GB)

心理学+NLP 教练技术+精神分析+亲子家庭教育+催眠+++等等 (更新超 2000GB)

大学-已更新至 9333 个课程+高中+初中+小学-全套资料 (超过 2 万 GB)

陈安之	曾仕强	马云	杜云生	翟鸿燊	刘一秒	俞凌雄
王健林	余世维	雷军	周文强	安东尼罗宾	董明珠	李嘉诚
徐鹤宁	冯晓强	李践	刘克亚	罗伯特清崎	戴志强	李伟贤
苏引华	史玉柱	李强	俞敏洪	杰亚伯拉罕	周鸿祎	唐骏
梁凯恩	陈永亮	傅佩荣	贾长松	易发久	李彦宏	湖畔大学
李开复	慕泉	悟空道场	魏星	姬剑晶	其他名师全集	其他资料下载
王兴	王智华	智多星	陈文强	周导		微信 2270291360
泡妞	撩汉	泡仔	房中术	性福课		泡妞撩汉性福合集

注: 太多了, 无法全部一一列出。。。

全套专题系列【微信 1131084518】

记忆力训练	形象礼仪	健康养生	企业管理	沟通技巧
演讲与口才	经理修炼	MBA 讲座	时间管理	战略经营
企业文化	销售心理	管理素质	国学讲座	执行力
团队管理	领导艺术	员工激励	潜能激发	谈判技巧
绩效管理	薪酬管理	43 份直销制度	电话销售	人力管理
客户服务	创业指南	市场营销	餐饮管理	保险讲座
品牌营销	酒店管理	汽车 4S 店	众筹资料	销售技巧

兴趣爱好:	钓鱼教程	魔术教学	炒股教学	美术教学	书法教学
音乐乐器:	萨克斯教学	电子琴教学	小提琴	古筝教学	钢琴教学
	吉他教学				
体育运动:	篮球教学	足球教学	羽毛球教学	乒乓球教学	太极拳教学
	围棋教学	高尔夫球			
生活实用:	插花教学	茶艺-茶道	唱歌教学	单反相机摄影	毛线编织
	小吃+美食				
语言学习:	英语				
电脑 IT:	办公 office	PS 美工教学			

暗号: 666

免费领取资料微信

1131084518

微信1131084518

撩汉liaohan.net

最好资源zuihaoziyuan.com



	0.5 TFLOPS					
	1.3TFLOPS					
14 21TOPS						
32TOPS	16TFLOPS					
19.3TOPS	9.6TFLOPS					
16TOPS	8TFLOPS					12nm FFC
3TOPS	100GFLOPS	300GOPS(INT16)				22nm
3TOPS	0.3T	0.1T				28nm
NA	4.8T	NA				T16
/	/	/				16 nm UltraScale™
/	/	/				16 nm UltraScale™
320 TOPS						12nm
30 TOPS						12nm
支持	支持	不支持				22nm
3.2TOPS	0.5TOPS					12nm
1TOPS						T28
2TOPS						T12
4TOPS						T12
0.5TOPS						T28
1TOPS						T28
1.5TOPS						T12
4TOPS						T12
1.2TOPS@600MHz						T C28HPC+
						N7+
√	√	√				7nm
√	√	√				55nm
16GOPS	2GOPS	2GOPS				40nm
12.8GOPS@100MHZ						T 40nm
0.2 - 4TOPS	n/a	n/a				28nm